Mathematics of Information & Entropy

Volume I: Foundations of Information Theory

Shannon entropy, mutual information, channel capacity, the asymptotic equipartition property, and source coding

---

Every derivation step by step

Every symbol defined

Explicit numerical examples

Geometric and combinatorial intuitions

Silvano Salvador

Mathematics of Information & Entropy Series — Volume I of III

Contents

1.1 Shannon Entropy: Definition and Axioms 4

1.2 Joint Entropy, Conditional Entropy, and the Chain Rule 6

1.3 Jensen's Inequality and the Log-Sum Inequality 9

1.4 Entropy of Continuous Random Variables 10

1.5 Entropy Rate of Stochastic Processes 12

1.6 Rényi and Tsallis Entropy: Generalized Information Measures 13

2.1 Mutual Information 21

2.2 KL Divergence and Its Properties 22

2.3 The Data-Processing Inequality 23

2.4 Pinsker's Inequality and Total Variation 24

2.5 f-Divergences and Variational Representations 25

2.6 Maximum Entropy, Exponential Families, and Lagrange Duality 30

3.1 Typical Sequences and the AEP 34

3.2 Source Coding: Definitions and Kraft's Inequality 35

3.3 Shannon's Source Coding Theorem 36

3.4 Entropy and Statistical Mechanics: A Bridge 37

3.5 Universal Source Coding and the LZ Algorithm 38

3.6 The Method of Types in Detail 38

3.7 Hypothesis Testing, Stein's Lemma, and the Chernoff Information 43

4.1 Discrete Memoryless Channels 48

4.2 Shannon's Channel Coding Theorem 49

4.3 The Gaussian Channel and Water-Filling 50

4.4 The Blahut–Arimoto Algorithm 51

4.5 Strong Converse and the Channel Dispersion 51

4.6 Channels with Memory, Feedback, and the Information Spectrum Method 53

4.7 Joint Source-Channel Coding and the Separation Theorem 57

Numerical Tour: Part 1 — From Distributions to Entropy Rates 61

Formula Reference: Part 1 — Entropy 62

Numerical Tour: Part 2 — Mutual Information, KL Divergence, and the Data-Processing Inequality 64

Formula Reference: Part 2 — Mutual Information and Divergence 65

Numerical Tour: Part 3 — AEP, Typical Sets, and Source Codes 66

Formula Reference: Part 3 — AEP and Source Coding 67

Numerical Tour: Part 4 — Channel Capacity via Blahut–Arimoto and Water-Filling 68

Formula Reference: Part 4 — Channel Capacity 69

Computational Notes: Algorithms and Implementation 71

Glossary of Key Terms and Symbols 75

Open Problems and Research Frontiers 79

A Measure-Theoretic Foundations of Entropy 82

B Convex Duality, Legendre Transform, and Entropy 83

C Entropy Bounds in Combinatorics 83

D Entropy and Hypothesis Testing: Sanov's Theorem 84

E Preview of Network Information Theory 85

F Proofs of Selected Theorems 87

G Statistical and Reference Tables 90

H Notation Index 92

I Quick Reference: Generalized Entropies, Divergences, and Operational Exponents 93

J A Catalogue of Exponential Families 96

Extended Case Study: Information-Theoretic Design of a Deep-Space Communication Link 98

Cross-Volume Connections and Methodological Synthesis 101

Preface to Volume I

Scope and motivation. Information theory is one of the most elegant mathematical edifices of the twentieth century. Born in a single paper by Claude Shannon in 1948, it established for the first time a rigorous, quantitative theory of communication: how much information a source produces, how efficiently it can be compressed, and how reliably it can be transmitted through a noisy channel. These three fundamental questions — the content, the compression, and the capacity — define the intellectual arc of this volume.

The central concept is *entropy* $H(X)$, a non-negative functional of a probability distribution that measures uncertainty or surprise. Everything else in classical information theory is built from it: mutual information quantifies the statistical dependence between two random variables; channel capacity is the supremum of mutual information over all input distributions; source coding theorems show that entropy is the fundamental limit of lossless data compression. Understanding entropy deeply — its axiomatic characterisation, its operational meaning, its connections to physics and to statistical inference — is the intellectual core of this volume.

Structure. The volume is organised into four parts, designed to be read sequentially. Each part begins with the foundational sections outlined below and closes with one or two extension sections that generalize the core ideas, develop their variational and dual representations, and connect them forward to later volumes of the series.

- **Part 1: Entropy and Information.** The axiomatic and probabilistic foundations: entropy for discrete and continuous random variables, the uniqueness theorem (Shannon–Khinchin axioms), joint entropy, conditional entropy, chain rules, Jensen's inequality as the unifying analytic tool, the information inequality $H(X) \leq \log |\mathcal{X}|$, and the entropy rate of stochastic processes; the part closes with the Rényi and Tsallis generalizations of entropy and their associated divergences (Section 1.6).

- **Part 2: Mutual Information and Divergence.** Mutual information as the reduction in uncertainty due to observation, relative entropy (KL divergence) as an asymmetric measure of distributional discrepancy, the data-processing inequality, sufficient statistics, and Pinsker's inequality connecting total variation to KL divergence; the part closes with the general f-divergence framework and its variational representations (Section 2.5), and with maximum entropy, exponential families, and the Bregman/Legendre duality between entropy and the log-partition function (Section 2.6).

- **Part 3: The Asymptotic Equipartition Property.** The law of large numbers for information: the AEP theorem and its proof, typical sets and their properties, the equipartition of probability among typical sequences, source coding via typical sets, and the relationship between AEP and thermodynamic entropy; the part closes with the method of types and Sanov's theorem (Section 3.6) and with the hypothesis-testing exponents — Stein's lemma and the Chernoff information — that the method of types makes precise (Section 3.7).

- **Part 4: Channel Capacity and Coding Theorems.** Discrete memoryless channels, the channel capacity as a max-min problem, Shannon's channel coding theorem (achievability and converse), the binary symmetric channel and binary erasure channel as canonical examples, the Gaussian channel and the water-filling formula, and the Blahut–Arimoto algorithm for computing capacity; the part closes with channels with memory, feedback, and the information spectrum method (Section 4.6), and with the joint source-channel coding (separation) theorem that ties the entropy-rate and capacity threads of the volume together (Section 4.7).

Mathematical prerequisites. The reader is assumed to be familiar with probability theory at the graduate level (random variables, expectations, convergence theorems), real analysis (continuity, compactness, convexity), and basic combinatorics. Measure-theoretic probability is used where needed but introduced as required.

Notation conventions. Throughout this volume:

- X, Y, Z denote discrete random variables taking values in finite alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$.
- $p(x) = \Pr(X = x)$, $p(x, y) = \Pr(X = x, Y = y)$, $p(y|x) = \Pr(Y = y|X = x)$.
- $H(X)$: Shannon entropy of X.
- $H(X|Y)$: conditional entropy of X given Y.
- $I(X;Y)$: mutual information between X and Y.
- $D_{\mathrm{KL}}(p \,\|\, q)$: KL divergence from q to p.
- log denotes logarithm base 2 (entropy in bits) unless stated otherwise. Natural logarithm gives entropy in nats; the factor $\ln 2$ converts.
- $h(p) = -p \log p - (1-p) \log(1-p)$: binary entropy function.
- $[n] = \{1, 2, \ldots, n\}$.
- $X^n = (X_1, X_2, \ldots, X_n)$: a sequence of n random variables.

Each section follows a consistent structure: (1) a formal **Definition Box** establishing mathematical objects; (2) rigorous derivations with no steps omitted; (3) at least one **Example Box** with explicit numerical computations; (4) extensions and connections to adjacent theory.

PART 1

Entropy and Information

Axiomatic and probabilistic foundations: Shannon entropy, the uniqueness theorem, joint and conditional entropy, chain rules, and Jensen's inequality.

1.1 Shannon Entropy: Definition and Axioms

Definition: Shannon Entropy

Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \Pr(X = x)$. The *Shannon entropy* of X is

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

with the convention $0 \log 0 = 0$ (justified by $\lim_{p \to 0^+} p \log p = 0$). When logarithms are base 2, $H(X)$ is measured in *bits*; in base e, in *nats*.

Combinatorial motivation. Suppose we draw n i.i.d. samples $X_1, X_2, \dots, X_n$ from p. By the law of large numbers, symbol x appears approximately $np(x)$ times. The number of such sequences is

$$\binom{n}{np(x_1), np(x_2), \dots} = \frac{n!}{\prod_x (np(x))!}.$$

By Stirling's approximation $\ln k! \approx k \ln k - k$, the log of this multinomial coefficient is approximately

$$n \ln n - n - \sum_x [np(x) \ln(np(x)) - np(x)] = -n \sum_x p(x) \ln p(x) = n H_e(X),$$

where H_e is entropy in nats. Thus $e^{nH_e(X)}$ counts the number of *typical* sequences, a combinatorial interpretation central to Part 3.

Properties of entropy.

1. **Non-negativity**: $H(X) \geq 0$, with equality iff X is deterministic ($\exists\, x_0 : p(x_0) = 1$).
2. **Symmetry**: $H(X)$ depends only on the probability values, not on the labels of the outcomes.
3. **Maximality**: For fixed $|\mathcal{X}|$,
$$H(X) \leq \log |\mathcal{X}|,$$
with equality iff X is uniformly distributed on $\mathcal{X}$.
4. **Continuity**: $H(X)$ is a continuous function of the probabilities $(p(x))_{x \in \mathcal{X}}$.
5. **Expansibility**: Adding a symbol of probability zero does not change $H(X)$.

Proof of maximality via Jensen's inequality. For the concave function $\phi(t) = -t \log t$ (equivalently, use $\phi(t) = \log(1/t)$):

$$\begin{aligned} H(X) &= \sum_x p(x) \log \frac{1}{p(x)} \\ &= \sum_x p(x) \log \frac{1}{p(x)} \cdot \frac{|\mathcal{X}|}{|\mathcal{X}|} \leq \log \sum_x p(x) \cdot \frac{1}{p(x)} \cdot \frac{1}{|\mathcal{X}|} \cdot |\mathcal{X}| = \log |\mathcal{X}|. \end{aligned} \tag{1}$$

More directly: by the log-sum inequality or Jensen's applied to log:

$$H(X) = \sum_x p(x) \log \frac{1}{p(x)} \leq \log\left(\sum_x p(x) \cdot \frac{1}{p(x)}\right) = \log|\mathcal{X}|,$$

with equality iff $1/p(x) = |\mathcal{X}|$ for all x, i.e., p is uniform.

Uniqueness Theorem (Shannon–Khinchin Axioms)

The Shannon entropy H is the *unique* function (up to a positive multiplicative constant) satisfying:

1. **Continuity**: $H(p_1, \dots, p_k)$ is continuous in all arguments.
2. **Maximality**: $H(p_1, \dots, p_k) \leq H(1/k, \dots, 1/k)$.
3. **Expansibility**: $H(p_1, \dots, p_k, 0) = H(p_1, \dots, p_k)$.
4. **Additivity (chain rule)**: For a compound experiment with a two-stage structure:
$$H(p_{11}, \dots, p_{km}) = H(p_1, \dots, p_k) + \sum_{i=1}^{k} p_i H\left(\frac{p_{i1}}{p_i}, \dots, \frac{p_{im}}{p_i}\right),$$
where $p_i = \sum_j p_{ij}$.

Proof sketch of uniqueness. Let $H_n = H(1/n, \dots, 1/n)$ for a uniform distribution on n symbols. From axiom 4 applied to a two-stage experiment (first choose among n equal groups, then among m equal elements within each group), one derives $H_{nm} = H_n + H_m$, so $H_n = C \log n$ for some constant $C > 0$. The general result then follows by approximating any rational distribution by a refinement of a uniform distribution and using continuity.

Entropy as expected surprise. Define the *information content* (or self-information) of outcome x as $\iota(x) = \log(1/p(x))$. Then $H(X) = \mathbb{E}[\iota(X)]$: entropy is the expected surprise. The function $\iota(x)$ satisfies:

- $\iota(x) \geq 0$ (observing an event gives non-negative information).
- $\iota(x) = 0$ iff $p(x) = 1$ (a certain event carries no information).
- For independent events $A \perp B$: $\iota(A \cap B) = \iota(A) + \iota(B)$ (information is additive for independent observations).

These three properties, combined with monotonicity in p, uniquely determine $\iota(x) = -\log p(x)$.

Example — Entropy of a biased coin

Let $X \in \{0, 1\}$ with $\Pr(X = 1) = p$ and $\Pr(X = 0) = 1 - p$. The binary entropy function is

$$h(p) = H(X) = -p \log p - (1-p) \log(1-p).$$

Key values:

- $h(0) = h(1) = 0$ (deterministic, no uncertainty).
- $h(1/2) = 1$ bits (maximum uncertainty; fair coin).
- $h(0.1) = -0.1 \log_2(0.1) - 0.9 \log_2(0.9) = 0.332 + 0.152 = 0.469$ bits.
- $h(0.3) = -0.3 \log_2(0.3) - 0.7 \log_2(0.7) = 0.521 + 0.360 = 0.881$ bits.

The function $h(p)$ is concave, achieves its maximum 1 bit at $p = 1/2$, and is symmetric around $p = 1/2$. Its derivative is $h'(p) = \log[(1-p)/p]$, which equals zero at $p = 1/2$.

Entropy of a die. A fair six-sided die has $H(X) = \log_2 6 = 2.585$ bits. An unfair die with face probabilities $(1/2, 1/4, 1/8, 1/16, 1/32, 1/32)$ has

$$H = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{16}\cdot 4 + \frac{1}{32}\cdot 5 + \frac{1}{32}\cdot 5 = 1.9375 \text{ bits}.$$

The fair die has higher entropy, as expected.

Concavity of entropy. H is a concave function of the probability vector $\mathbf{p} = (p_1, \dots, p_k)$: for any two distributions $\mathbf{p}, \mathbf{q}$ and $\lambda \in [0,1]$,

$$H(\lambda \mathbf{p} + (1-\lambda)\mathbf{q}) \geq \lambda H(\mathbf{p}) + (1-\lambda) H(\mathbf{q}).$$

Proof. The function $f(t) = -t \log t$ is strictly concave on $[0,1]$ (since $f''(t) = -1/(t \ln 2) < 0$). Therefore

$$-(\lambda p_i + (1-\lambda)q_i)\log(\lambda p_i + (1-\lambda)q_i) \geq \lambda(-p_i \log p_i) + (1-\lambda)(-q_i \log q_i),$$

and summing over i gives the result. This concavity has operational meaning: mixing two sources cannot decrease entropy.

Rényi entropy. The Shannon entropy is a member of the *Rényi family*:

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \sum_x p(x)^\alpha, \quad \alpha \geq 0,\ \alpha \neq 1.$$

As $\alpha \to 1$, $H_\alpha(X) \to H(X)$ (by L'Hôpital). Special cases: $H_0 = \log |\mathrm{supp}(p)|$ (Hartley entropy), $H_2 = -\log \sum_x p(x)^2$ (collision entropy), $H_\infty = -\log \max_x p(x)$ (min-entropy). All Rényi entropies are non-increasing in α: $H_\alpha \geq H_\beta$ for $\alpha \leq \beta$.

Example — Rényi entropies of a geometric distribution

Let $X \sim \mathrm{Geom}(1/2)$: $p(k) = (1/2)^k$ for $k = 1, 2, 3, \dots$ (truncated to first 8 values for numerical illustration, renormalized: $p(k) \propto (1/2)^k$, so $p(k) = (1/2)^k/(1 - 1/2^8) \approx (1/2)^k \cdot 256/255$).
For the exact geometric with $p(k) = (1-q)q^{k-1}$, $q = 1/2$:

$$H_\alpha = \frac{1}{1-\alpha} \log \left[\frac{(1-q)^\alpha}{1-q^\alpha} \right].$$

At $q = 1/2$: $H_1 = \log_2(2)/(1/2) = 2$ bits. More precisely, $H(X) = -\sum_{k=1}^\infty (1/2)^k \log_2 (1/2)^k = \sum_{k=1}^\infty k(1/2)^k = 2$ bits. Collision entropy: $H_2 = -\log_2 \sum_{k=1}^\infty (1/4)^k = -\log_2(1/3) = \log_2 3 \approx 1.585$ bits. Min-entropy: $H_\infty = -\log_2(1/2) = 1$ bits. Indeed $H_\infty \leq H_2 \leq H_1$, confirming the monotonicity of Rényi entropies.

Entropy and thermodynamics. Boltzmann's entropy $S = k_B \ln W$ (where W is the number of microstates) is related to Shannon entropy by $S = k_B \ln 2 \cdot H$ when all microstates are equally probable. The second law of thermodynamics — entropy non-decreases in isolated systems — has an information-theoretic counterpart: the data-processing inequality (Section 2.4). Jaynes' maximum entropy principle provides a formal bridge: the equilibrium distribution in statistical mechanics is the one that maximises Shannon entropy subject to the macroscopic constraints (energy, volume, etc.).

1.2 Joint Entropy, Conditional Entropy, and the Chain Rule

Joint and Conditional Entropy

For random variables X and Y with joint distribution $p(x, y)$:

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \quad \text{(joint entropy)}, \tag{2}$$

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) = \sum_{x} p(x) H(Y|X=x) \quad \text{(conditional entropy)}. \tag{3}$$

Chain rule for entropy.

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y).$$

Proof.

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\ &= -\sum_{x,y} p(x,y) \log[p(x)p(y|x)] \\ &= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y|x) \\ &= H(X) + H(Y|X). \end{aligned}$$

General chain rule. For n random variables $X_1, X_2, \dots, X_n$:

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} H(X_i \,|\, X_1, \dots, X_{i-1}).$$

By induction on n, applying the two-variable chain rule at each step.

Conditioning reduces entropy.

$$H(X|Y) \le H(X),$$

with equality iff $X \perp Y$. This formalises the intuition that additional information cannot increase uncertainty.

Proof.

$$\begin{aligned} H(X) - H(X|Y) &= -\sum_{x} p(x) \log p(x) + \sum_{x,y} p(x,y) \log p(x|y) \\ &= \sum_{x,y} p(x,y) \log \frac{p(x|y)}{p(x)} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \ge 0, \end{aligned}$$

by the non-negativity of KL divergence (Gibbs' inequality, proved in Section 2.2). Equality holds iff $p(x|y) = p(x)$ for all x, y, i.e., $X \perp Y$.

Sub-additivity.

$$H(X_1, \dots, X_n) \le \sum_{i=1}^{n} H(X_i),$$

with equality iff $X_1, \dots, X_n$ are mutually independent. From the chain rule and the fact that $H(X_i \,|\, X_1, \dots, X_{i-1}) \le H(X_i)$.

Example — Joint entropy of two fair coins

Let $X, Y \in \{0,1\}$ be i.i.d. Bernoulli$(1/2)$. Then $p(x,y) = 1/4$ for all (x,y).

$$H(X,Y) = -4 \cdot \frac{1}{4}\log_2 \frac{1}{4} = 2 \text{ bits} = H(X) + H(Y).$$

Independence gives equality in sub-additivity. Now let $Y = X$ (perfectly correlated): $p(0,0) = p(1,1) = 1/2$, $p(0,1) = p(1,0) = 0$.

$$H(X,Y) = -2 \cdot \frac{1}{2}\log \frac{1}{2} = 1 \text{ bits}, \quad H(Y|X) = 0.$$

The chain rule: $H(X,Y) = H(X) + H(Y|X) = 1 + 0 = 1$ bits. Conditioning on X eliminates all uncertainty in Y.

Fano's inequality. Let X and $\hat{X}$ be random variables on the same alphabet $\mathcal{X}$, and define the probability of error $P_e = \Pr(\hat{X} \neq X)$. Then

$$H(X|\hat{X}) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1),$$

where $h(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function.

Proof. Define the error indicator $E = \mathbf{1}[\hat{X} \neq X]$, so $\Pr(E=0) = 1 - P_e$ and $\Pr(E=1) = P_e$. By the chain rule:

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X,\hat{X}) = H(X|\hat{X}) + 0, \\ H(E, X|\hat{X}) &= H(E|\hat{X}) + H(X|E,\hat{X}) \leq H(E) + H(X|E,\hat{X}). \end{aligned}$$

When $E = 0$, $X = \hat{X}$, so $H(X|E=0,\hat{X}) = 0$. When $E = 1$, X ranges over $\mathcal{X} \setminus \{\hat{X}\}$, giving $H(X|E=1,\hat{X}) \leq \log(|\mathcal{X}| - 1)$. Combining:

$$H(X|\hat{X}) \leq H(E) + P_e \log(|\mathcal{X}| - 1) \leq h(P_e) + P_e \log(|\mathcal{X}| - 1).$$

Fano's inequality is fundamental for proving converse coding theorems: if the channel capacity is C and we attempt to communicate at rate $R > C$, then P_e is bounded below, showing that reliable communication is impossible.

Han's inequality. For any n random variables $X_1, \ldots, X_n$:

$$H(X_1, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n).$$

This generalises sub-additivity and is used in the proof of the blowing-up lemma in multi-user information theory.

Example — Conditional entropy and weather prediction

Suppose tomorrow's weather $Y \in \{\text{sun}, \text{rain}, \text{cloud}\}$ has marginal distribution $p_Y = (0.5, 0.3, 0.2)$, giving $H(Y) = -0.5\log_2 0.5 - 0.3\log_2 0.3 - 0.2\log_2 0.2 = 0.5 + 0.521 + 0.464 = 1.485$ bits.

Suppose today's weather $X \in \{\text{sun}, \text{rain}\}$ provides information:

$$\begin{aligned} H(Y|X=\text{sun}) &= h(0.7, 0.1, 0.2) = -0.7\log 0.7 - 0.1\log 0.1 - 0.2\log 0.2 \\ &= 0.357 + 0.332 + 0.464 = 1.153 \text{ bits}, \\ H(Y|X=\text{rain}) &= h(0.1, 0.7, 0.2) = 1.153 \text{ bits} \quad \text{(symmetric)}. \end{aligned}$$

With $\Pr(X=\text{sun}) = 0.6$ and $\Pr(X=\text{rain}) = 0.4$:

$$H(Y|X) = 0.6 \times 1.153 + 0.4 \times 1.153 = 1.153 \text{ bits} < H(Y) = 1.485 \text{ bits}.$$

Knowing today's weather reduces entropy by $1.485 - 1.153 = 0.332$ bits. This reduction equals the mutual information $I(X;Y) = 0.332$ bits (Section 2.1).

1.3 Jensen's Inequality and the Log-Sum Inequality

Jensen's Inequality

Let $\phi : \mathcal{I} \to \mathbb{R}$ be a convex function on an interval $\mathcal{I}$, and let X be a random variable taking values in $\mathcal{I}$ with finite expectation. Then

$$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)].$$

Equality holds iff X is constant almost surely (strict convexity case).

Proof. By the supporting hyperplane theorem, for convex ϕ there exists a linear function $\ell(x) = ax + b$ (a supporting line at $\mu = \mathbb{E}[X]$) such that $\ell(\mu) = \phi(\mu)$ and $\ell(x) \leq \phi(x)$ for all x. Taking expectations: $\phi(\mu) = \ell(\mu) = \mathbb{E}[\ell(X)] = a\mathbb{E}[X] + b \leq \mathbb{E}[\phi(X)]$.

Concave case. For concave ϕ: $\phi(\mathbb{E}[X]) \geq \mathbb{E}[\phi(X)]$. Since $\phi(t) = \log t$ is strictly concave,

$$\log(\mathbb{E}[X]) \geq \mathbb{E}[\log X],$$

with equality iff X is constant. This is the fundamental analytic engine behind nearly every inequality in information theory.

Log-Sum Inequality

For non-negative numbers $a_1, \ldots, a_n$ and positive numbers $b_1, \ldots, b_n$:

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i},$$

with equality iff a_i/b_i is constant for all i.

Proof. Let $A = \sum_i a_i$ and $B = \sum_i b_i$. Write $a_i = A\alpha_i$ and $b_i = B\beta_i$ where $\alpha_i = a_i/A$ and $\beta_i = b_i/B$ are probability distributions. Then

$$\sum_i a_i \log \frac{a_i}{b_i} = A \sum_i \alpha_i \log \frac{A\alpha_i}{B\beta_i} = A \log \frac{A}{B} + A \sum_i \alpha_i \log \frac{\alpha_i}{\beta_i} = A \log \frac{A}{B} + A \cdot D_{\mathrm{KL}}(\alpha \,\|\, \beta)\,.$$

Since $D_{\mathrm{KL}}(\alpha \,\|\, \beta) \geq 0$, the result follows.

Applications to entropy inequalities. The log-sum inequality immediately gives:

(*i*) *Information inequality:* $H(X) \le \log|\mathcal{X}|$. Take $a_i = p(x_i)$, $b_i = 1/|\mathcal{X}|$:

$$\sum_i p(x_i) \log \frac{p(x_i)}{1/|\mathcal{X}|} = -H(X) + \log|\mathcal{X}| \ge 0.$$

(*ii*) *Non-negativity of KL divergence:* $D_{\mathrm{KL}}(p \,\|\, q) \ge 0$. Take $a_i = p(x_i)$, $b_i = q(x_i)$, with $A = B = 1$:

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_i p_i \log \frac{p_i}{q_i} \ge \sum_i p_i \log \frac{1}{1} = 0.$$

(*iii*) *Concavity of entropy:* already proved in Section 1.1 via $f''(t) < 0$.

Example — Log-sum inequality applied to data fusion

Two sensors each report a distribution over three classes $\mathcal{Y} = \{A, B, C\}$: $p = (0.5, 0.3, 0.2)$ and $q = (0.2, 0.5, 0.3)$. The average distribution $r = (p + q)/2 = (0.35, 0.40, 0.25)$.
By concavity of entropy:

$$H(r) \ge \tfrac{1}{2}H(p) + \tfrac{1}{2}H(q).$$

Check: $H(p) = 1.485$ bits, $H(q) = -0.2\log 0.2 - 0.5\log 0.5 - 0.3\log 0.3 = 0.464 + 0.5 + 0.521 = 1.485$ bits (same entropy by symmetry), $H(r) = -0.35\log 0.35 - 0.40\log 0.40 - 0.25\log 0.25 = 0.530 + 0.529 + 0.500 = 1.559$ bits. Indeed $1.559 \ge 1.485$. Mixing distributions increases entropy: the fused distribution r is more uncertain than either individual sensor.

Gibbs' inequality. The log-sum inequality with $a_i = p_i$, $b_i = q_i$, $\sum a_i = \sum b_i = 1$ gives:

$$\sum_i p_i \log \frac{p_i}{q_i} \ge \left(\sum_i p_i\right) \log \frac{\sum_i p_i}{\sum_i q_i} = 0.$$

This is Gibbs' inequality: $D_{\mathrm{KL}}(p \,\|\, q) \ge 0$, the most important inequality in information theory.

Convexity/concavity of entropy and KL divergence.

- $H(p)$ is concave in p: proved via $f''(t) < 0$.
- $D_{\mathrm{KL}}(p \,\|\, q)$ is convex in the pair (p, q): for any $\lambda \in (0, 1)$ and pairs (p_1, q_1), (p_2, q_2):

  $$D_{\mathrm{KL}}(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2) \le \lambda D_{\mathrm{KL}}(p_1 \,\|\, q_1) + (1-\lambda) D_{\mathrm{KL}}(p_2 \,\|\, q_2)\,.$$

  This follows from the log-sum inequality.

1.4 Entropy of Continuous Random Variables

Differential Entropy

For a continuous random variable X with density $f(x)$ on $\mathbb{R}$ (or $\mathbb{R}^d$), the *differential entropy* is

$$h(X) = -\int f(x) \log f(x)\, dx,$$

provided the integral converges absolutely. When log is base e, the unit is nats.

Differential entropy is **not** the limit of discrete entropy: it can be negative, it is not invariant under change of coordinates, and it does not directly measure uncertainty in the same operational sense as discrete entropy. These differences are fundamental and must not be ignored.

Differential entropy of standard distributions.

Uniform on $[a, b]$: $f(x) = 1/(b-a)$.

$$h(X) = -\int_a^b \frac{1}{b-a} \log \frac{1}{b-a}\, dx = \log(b-a).$$

Negative when $b - a < 1$; positive when $b - a > 1$.

Gaussian $\mathcal{N}(\mu, \sigma^2)$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$.

$$\begin{aligned} h(X) &= -\mathbb{E}\left[\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{(X-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}[(X-\mu)^2] \\ &= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2}\log(2\pi e\sigma^2) \text{ nats.} \end{aligned}$$

In bits: $h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$.

Exponential $\mathrm{Exp}(\lambda)$: $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$.

$$h(X) = -\int_0^\infty \lambda e^{-\lambda x}(\ln\lambda - \lambda x)\, dx = -\ln\lambda + \lambda \cdot \frac{1}{\lambda} = 1 - \ln\lambda \text{ nats.}$$

Multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \square)$ on $\mathbb{R}^d$:

$$h(\mathbf{X}) = \frac{1}{2}\operatorname{logdet}(2\pi e\, \square) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log\det \square \text{ nats.}$$

Maximum entropy for given moments. The Gaussian maximises differential entropy over all distributions with fixed mean and variance.

Maximum Entropy Theorem

Among all distributions on $\mathbb{R}$ with mean μ and variance σ^2, the Gaussian $\mathcal{N}(\mu, \sigma^2)$ uniquely maximises $h(X)$.

Proof. Let $g = \mathcal{N}(\mu, \sigma^2)$ and f any other density with $\mathbb{E}_f[X] = \mu$, $\mathbb{E}_f[X^2] = \mu^2 + \sigma^2$. Then

$$0 \leq D_{\mathrm{KL}}(f \,\|\, g) = \int f(x) \log \frac{f(x)}{g(x)}\, dx = -h(f) - \int f(x) \log g(x)\, dx.$$

Since $\log g(x) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$:

$$\int f(x) \log g(x)\, dx = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\sigma^2}{2\sigma^2} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2} = -h(g).$$

Thus $0 \leq D_{\mathrm{KL}}(f \,\|\, g) = -h(f) + h(g)$, giving $h(f) \leq h(g)$.

Analogously: over distributions on $[0, \infty)$ with fixed mean $1/\lambda$, the exponential $\mathrm{Exp}(\lambda)$ maximises entropy. Over distributions on $[a, b]$, the uniform $\mathrm{Unif}(a, b)$ maximises entropy.

Example — Differential entropy comparison across distributions

Fix mean $\mu = 0$ and variance $\sigma^2 = 1$. In nats:

$$h(\mathcal{N}(0,1)) = \tfrac{1}{2}\ln(2\pi e) = \tfrac{1}{2}(1.838 + 1) = 1.419 \text{ nats},$$
$$h(\text{Laplace}(0,b)) \text{ with } \sigma^2 = 2b^2 : b = 1/\sqrt{2},\ h = 1 + \ln(2/\sqrt{2}) = 1.347 \text{ nats},$$
$$h(\text{Logistic}(0,s)) \text{ with } \sigma^2 = \pi^2 s^2/3 : s = \sqrt{3}/\pi,\ h = 2\ln s + 2 \approx 1.373 \text{ nats}.$$

The Gaussian achieves the highest differential entropy among these three, confirming the maximum entropy theorem. The Laplace distribution has heavier tails but less entropy than the Gaussian at fixed variance.

Quantization and the relationship between discrete and differential entropy. If X is a continuous random variable, quantize it to resolution Δ: $X^\Delta = \lfloor X/\Delta \rfloor \cdot \Delta$. Then

$$H(X^\Delta) = h(X) - \log \Delta + O(\Delta),$$

showing that as $\Delta \to 0$, the discrete entropy $H(X^\Delta)$ grows like $-\log \Delta$ while the excess $H(X^\Delta) + \log \Delta \to h(X)$. The differential entropy is thus the finite part of the discrete entropy in the fine-quantization limit.

1.5 Entropy Rate of Stochastic Processes

Entropy Rate

The *entropy rate* of a stationary stochastic process $\{X_i\}_{i=1}^{\infty}$ is

$$\mathcal{H} = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \dots, X_n),$$

provided the limit exists, and equivalently

$$\mathcal{H} = \lim_{n\to\infty} H(X_n \mid X_1, \dots, X_{n-1}).$$

Existence for stationary processes. For a stationary process, the sequence $a_n = H(X_1, \dots, X_n)$ is sub-additive: $a_{n+m} \le a_n + a_m$ (by sub-additivity of entropy). Therefore a_n/n converges (Fekete's lemma). The equivalence of the two limits follows because $H(X_1, \dots, X_n)/n$ and $H(X_n|X_1, \dots, X_{n-1})$ both converge to the same value for stationary processes.

Entropy rate of i.i.d. process. If $X_i \overset{\text{i.i.d.}}{\sim} p$, then $H(X_1, \dots, X_n) = nH(X_1)$, so $\mathcal{H} = H(X_1)$.

Entropy rate of a Markov chain. Let $\{X_n\}$ be a stationary first-order Markov chain with transition matrix $P = (p_{ij})$ and stationary distribution π. Then

$$\mathcal{H} = -\sum_{i,j} \pi_i p_{ij} \log p_{ij} = \sum_i \pi_i H(X_{n+1}|X_n = i).$$

This is the entropy rate of the Markov chain: the expected entropy of the next state given the current state, averaged over the stationary distribution.

Example — Entropy rate of a two-state Markov chain

States $\{0, 1\}$, transition matrix $P = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$. Stationary distribution: $\pi_0 = \beta/(\alpha+\beta)$, $\pi_1 = \alpha/(\alpha+\beta)$.
Conditional entropies:

$$H(X_{n+1}|X_n = 0) = h(\alpha) = -\alpha \log \alpha - (1-\alpha)\log(1-\alpha),$$
$$H(X_{n+1}|X_n = 1) = h(\beta).$$

Entropy rate:

$$\mathcal{H} = \frac{\beta}{\alpha+\beta} h(\alpha) + \frac{\alpha}{\alpha+\beta} h(\beta).$$

Numerical instance: $\alpha = 0.1$ (rare transitions from 0), $\beta = 0.2$. $\pi_0 = 2/3$, $\pi_1 = 1/3$. $h(0.1) = 0.469$ bits, $h(0.2) = 0.722$ bits. $\mathcal{H} = \frac{2}{3}(0.469) + \frac{1}{3}(0.722) = 0.313 + 0.241 = 0.554$ bits per symbol. Compare to i.i.d. with $\Pr(X=0) = 2/3$: $H = h(2/3) = 0.918$ bits per symbol. The Markov structure (persistence) reduces the entropy rate substantially.

Entropy rate and predictability. A process with entropy rate $\mathcal{H} = 0$ is perfectly predictable from its past. A process with $\mathcal{H} = H(X_1)$ is i.i.d. (no structure). Natural language has been estimated to have an entropy rate of approximately 1–1.5 bits per character (compared to the maximum $\log_2 27 \approx 4.75$ bits for English letters including space), reflecting the high redundancy in text.

Entropy rate and compression. The fundamental theorem of source coding (Section 4.1 below and Part 3) states that the minimum achievable compression rate for a stationary ergodic source equals its entropy rate $\mathcal{H}$. Any lossless code must use at least $\mathcal{H}$ bits per symbol; Huffman and arithmetic codes approach this limit.

1.6 Rényi and Tsallis Entropy: Generalized Information Measures

Shannon entropy is the unique information measure satisfying additivity for independent systems together with a small set of natural continuity and grouping axioms, but it is far from the only functional that deserves to be called an "entropy." Once the additivity axiom is relaxed to allow a one-parameter family of measures, a much richer landscape opens up, indexed by a real parameter α (or, in a different parametrization, q). This section introduces the two most important such families — Rényi entropy and Tsallis entropy — derives their basic properties, situates Shannon entropy as a limiting case of each, and develops the associated divergence measures. These generalized entropies reappear throughout the later volumes of this series: Rényi divergences underlie the modern treatment of differential privacy and large-deviation refinements, α-connections built from Rényi divergence generate the dualistic geometry of statistical manifolds (Volume III), and Tsallis entropy is the natural entropy functional for systems with long-range correlations and heavy-tailed equilibrium distributions.

Rényi Entropy

For a discrete random variable X with distribution p on a finite alphabet $\mathcal{X}$, and for $\alpha \in (0,1) \cup (1,\infty)$, the *Rényi entropy of order* α is

$$H_\alpha(X) = \frac{1}{1-\alpha} \log \left(\sum_{x \in \mathcal{X}} p(x)^\alpha \right).$$

The cases $\alpha \in \{0, 1, \infty\}$ are defined by continuity (limits), giving rise to the Hartley entropy, the Shannon entropy, and the min-entropy respectively.

The limit $\alpha \to 1$: recovering Shannon entropy. Write $\sum_x p(x)^\alpha = \sum_x p(x) \cdot p(x)^{\alpha-1} = \sum_x p(x) e^{(\alpha-1)\log p(x)}$. As $\alpha \to 1$, expand the exponential to first order:

$$\sum_x p(x)^\alpha = \sum_x p(x) \left[1 + (\alpha-1)\log p(x) + O((\alpha-1)^2)\right] = 1 + (\alpha-1)\sum_x p(x)\log p(x) + O((\alpha-1)^2).$$

Hence

$$\log\left(\sum_x p(x)^\alpha\right) = (\alpha-1)\sum_x p(x)\log p(x) + O((\alpha-1)^2),$$

and dividing by $(1-\alpha)$ and taking the limit gives

$$\lim_{\alpha \to 1} H_\alpha(X) = -\sum_x p(x)\log p(x) = H(X).$$

Shannon entropy is thus the value of the Rényi family at the removable singularity $\alpha = 1$, obtained by L'Hôpital's rule applied to $\frac{\log \sum_x p(x)^\alpha}{1-\alpha}$ as $\alpha \to 1$, which is exactly the computation above.

The limit $\alpha \to 0$: Hartley entropy. As $\alpha \to 0^+$, $p(x)^\alpha \to \mathbf{1}\{p(x) > 0\}$ for every x with $p(x) \in (0,1]$ (since $p(x)^0 = 1$ for $p(x) > 0$), so $\sum_x p(x)^\alpha \to |\mathrm{supp}(p)|$, the size of the support of p. Therefore

$$H_0(X) = \log |\mathrm{supp}(p)|,$$

the *Hartley entropy* (also called the max-entropy or zeroth-order Rényi entropy), which depends only on which outcomes have positive probability, not on the actual probability values. This is the original 1928 measure of Hartley, predating Shannon's 1948 paper: it counts the number of possible messages on a logarithmic scale, with no weighting by likelihood.

The limit $\alpha \to \infty$: min-entropy. For $\alpha > 1$, write $p_{\max} = \max_x p(x)$. Then

$$\sum_x p(x)^\alpha = p_{\max}^\alpha \sum_x \left(\frac{p(x)}{p_{\max}}\right)^\alpha.$$

Each ratio $p(x)/p_{\max} \in [0,1]$, with at least one term equal to 1 (the maximizing x), so as $\alpha \to \infty$ every ratio strictly less than 1 is raised to a vanishing power and the sum converges to the number of maximizing outcomes, call it $k \geq 1$. Hence $\sum_x p(x)^\alpha \to k \cdot p_{\max}^\alpha$, and

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\left(k\, p_{\max}^\alpha\right) = \frac{1}{1-\alpha}\left(\log k + \alpha \log p_{\max}\right) \xrightarrow{\alpha\to\infty} -\log p_{\max}.$$

Therefore

$$H_\infty(X) = -\log p_{\max} = -\log \max_x p(x),$$

the *min-entropy*. This is the most conservative of all the Rényi entropies: it depends only on the single most likely outcome, ignoring the shape of the rest of the distribution entirely.

Min-entropy quantifies the best an adversary can do in a single guess, and is the central quantity in randomness extraction and unpredictability arguments in cryptography.

The special case $\alpha = 2$: collision entropy.

$$H_2(X) = -\log\left(\sum_x p(x)^2\right) = -\log \Pr(X = X'),$$

where X' is an independent copy of X, since $\sum_x p(x)^2 = \Pr(X = X')$ is exactly the *collision probability* — the probability that two independent draws coincide. Collision entropy is the workhorse of universal hashing: the leftover hash lemma bounds the statistical distance to uniform of a hashed source in terms of H_2 rather than H or H_∞, because collision probability composes simply under hashing.

Monotonicity of Rényi Entropy in α

For fixed p, the map $\alpha \mapsto H_\alpha(X)$ is non-increasing on $(0, \infty)$. In particular,

$$H_0(X) \geq H_\alpha(X) \geq H(X) \geq H_\beta(X) \geq H_\infty(X) \qquad \text{for } 0 < \alpha < 1 < \beta.$$

Proof of monotonicity. Define $f(\alpha) = (1-\alpha)H_\alpha(X) = \log\left(\sum_x p(x)^\alpha\right)$. Monotonicity of H_α is equivalent to a convexity-type statement about f. Differentiate twice with respect to α. Write $S(\alpha) = \sum_x p(x)^\alpha$ and $L(x) = \log p(x)$, so $S(\alpha) = \sum_x e^{\alpha L(x)}$. Then $f(\alpha) = \log S(\alpha)$,

$$f'(\alpha) = \frac{S'(\alpha)}{S(\alpha)} = \frac{\sum_x L(x)p(x)^\alpha}{\sum_x p(x)^\alpha} = \mathbb{E}_{Q_\alpha}[L(X)],$$

where Q_α is the *tilted distribution* $Q_\alpha(x) \propto p(x)^\alpha$. This is the expectation of $\log p(X)$ under the tilted law, and

$$f''(\alpha) = \frac{S''(\alpha)S(\alpha) - S'(\alpha)^2}{S(\alpha)^2} = \mathbb{E}_{Q_\alpha}[L(X)^2] - \left(\mathbb{E}_{Q_\alpha}[L(X)]\right)^2 = \mathrm{Var}_{Q_\alpha}(L(X)) \geq 0.$$

So f is convex in α. Now consider $g(\alpha) = \frac{f(\alpha)}{1-\alpha}$ for $\alpha \neq 1$; a direct (if slightly lengthy) computation using the convexity of f together with $f(1) = 0$ (since $S(1) = \sum_x p(x) = 1$) shows that g is non-increasing. The cleanest route is via the power-mean inequality: for $\alpha < \beta$, Hölder's inequality applied to the vectors $(p(x)^\alpha)_x$ gives

$$\left(\sum_x p(x)^\beta\right)^{1/\beta} \leq \left(\sum_x p(x)^\alpha\right)^{1/\alpha} \qquad \text{for } 0 < \alpha < \beta,$$

i.e. the ℓ_α quasi-norm of the probability vector dominates its ℓ_β quasi-norm for $\alpha < \beta$ — a restatement of the fact that ℓ^p norms of a fixed vector decrease in p. Taking logarithms and rearranging the constants $\frac{1}{1-\alpha}$ and $\frac{1}{1-\beta}$ (which flip sign as α, β cross 1, but in a way that exactly compensates the inequality direction) yields $H_\alpha(X) \geq H_\beta(X)$.

Interpretation: bias toward the dominant outcomes. Increasing α increases the relative weight of large probabilities inside the sum $\sum_x p(x)^\alpha$ (since $p(x)^\alpha$ for $p(x)$ close to 1 shrinks more slowly than for $p(x)$ close to 0, when α grows). Consequently, high-α Rényi entropies are dominated by the most likely outcomes, while low-α Rényi entropies give more weight to the support size as a whole. Shannon entropy, at $\alpha = 1$, is the unique point of the family that weights $\log p(x)$ exactly by $p(x)$ itself — neither systematically over- nor under-weighting rare events relative to their own probability.

Example — The Rényi entropy spectrum of a four-outcome distribution

Let $p = (0.50,\ 0.25,\ 0.15,\ 0.10)$ on $\mathcal{X} = \{1, 2, 3, 4\}$ (logarithms in base 2).
$\alpha = 0$ **(Hartley):** all four probabilities are positive, so $H_0 = \log_2 4 = 2$ bits.
$\alpha = 1/2$**:** $\sum_x p(x)^{1/2} = \sqrt{0.50} + \sqrt{0.25} + \sqrt{0.15} + \sqrt{0.10} = 0.7071 + 0.5000 + 0.3873 + 0.3162 = 1.9106$.

$$H_{1/2} = \frac{1}{1 - 1/2} \log_2(1.9106) = 2 \log_2(1.9106) = 2(0.9341) = 1.868 \text{ bits}.$$

$\alpha = 1$ **(Shannon):** $H(X) = -\sum_x p(x) \log_2 p(x) = -(0.50 \log_2 0.50 + 0.25 \log_2 0.25 + 0.15 \log_2 0.15 + 0.10 \log_2 0.10)$. Term by term: $0.50 \times 1 = 0.500$; $0.25 \times 2 = 0.500$; $0.15 \times 2.737 = 0.4106$; $0.10 \times 3.322 = 0.3322$. Sum $= 1.7428$, so $H(X) = 1.743$ bits.
$\alpha = 2$ **(collision):** $\sum_x p(x)^2 = 0.25 + 0.0625 + 0.0225 + 0.01 = 0.3450$. $H_2 = \frac{1}{1-2} \log_2(0.345) = -\log_2(0.345) = 1.535$ bits.
$\alpha \to \infty$ **(min-entropy):** $p_{\max} = 0.50$, so $H_\infty = -\log_2 0.50 = 1$.
Summary of the spectrum:

$$H_0 = 2.000 > H_{1/2} = 1.868 > H_1 = 1.743 > H_2 = 1.535 > H_\infty = 1.000 \quad \text{(bits)},$$

exhibiting the monotone decrease guaranteed by the theorem above. Note that the gap between consecutive values shrinks as α grows past 1: the family "saturates" toward H_∞ rather than diverging.

Rényi divergence. The Rényi entropy of order α has a two-distribution analogue that generalizes KL divergence in exactly the same way that H_α generalizes H.

Rényi Divergence

For probability distributions p, q on $\mathcal{X}$ with $p \ll q$ (absolute continuity) and $\alpha \in (0, 1) \cup (1, \infty)$, the *Rényi divergence of order* α is

$$D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \left(\sum_x p(x)^\alpha q(x)^{1-\alpha} \right).$$

Recovering KL divergence. By the same L'Hôpital computation as before (now in the variable $\alpha - 1$ rather than $1 - \alpha$, which is why the sign convention in the definition flips relative to H_α),

$$\lim_{\alpha \to 1} D_\alpha(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = D_{\mathrm{KL}}(p \,\|\, q).$$

At $\alpha = 1/2$, $D_{1/2}(p\|q) = -2 \log \left(\sum_x \sqrt{p(x)q(x)} \right)$, and the inner sum $\sum_x \sqrt{p(x)q(x)}$ is the *Bhattacharyya coefficient* $BC(p, q) \in [0, 1]$; thus $D_{1/2}$ is, up to the factor -2, the log of the Bhattacharyya coefficient, and is closely related to the Hellinger distance via $H^2(p, q) = 1 - BC(p, q)$.

Monotonicity and non-negativity. $D_\alpha(p\|q) \geq 0$ for all $\alpha > 0$ with equality iff $p = q$, and $\alpha \mapsto D_\alpha(p\|q)$ is non-decreasing on $(0, \infty)$ (with the same convention that $\alpha = 1$ is filled in by continuity). This is the divergence analogue of the entropy monotonicity theorem, and it has an important consequence: *D_α for small α gives a lower bound on $D_{\mathrm{KL}}(p \,\|\, q)$*, which is the basis for many minimax lower bounds in statistics, since D_α for $\alpha < 1$ can sometimes be computed in closed form (e.g. for exponential families) when $D_{\mathrm{KL}}(p \,\|\, q)$ cannot.

Data-processing for Rényi divergence. For $\alpha \in (0, 1)$, D_α satisfies the same data-processing

inequality as KL divergence: if $X \to Y$ via a channel (Markov kernel) applied identically under both p and q, then $D_\alpha(p_Y \| q_Y) \le D_\alpha(p_X \| q_X)$. For $\alpha > 1$ the inequality continues to hold, but the proof requires more care (it follows from the joint convexity of $(p,q) \mapsto \sum_x p(x)^\alpha q(x)^{1-\alpha}$ in (p,q) for $\alpha \in (0,1)$, and from a duality argument for $\alpha > 1$). The upshot is that *every* Rényi divergence is a valid "distinguishability" measure compatible with data processing — a property that is essential for their use in differential privacy, where D_α bounds the privacy loss of a randomized mechanism under post-processing.

Example — Rényi divergence between two Bernoulli distributions

Let $p = \text{Bernoulli}(0.7)$, $q = \text{Bernoulli}(0.5)$, base-2 logarithms.
KL divergence ($\alpha \to 1$):

$$D_{\mathrm{KL}}(p \,\|\, q) = 0.7 \log_2 \frac{0.7}{0.5} + 0.3 \log_2 \frac{0.3}{0.5} = 0.7(0.485) + 0.3(-0.737) = 0.3395 - 0.2211 = 0.1184 \text{ bits.}$$

Rényi divergence of order $\alpha = 2$:

$$\sum_x p(x)^2 q(x)^{-1} = \frac{0.7^2}{0.5} + \frac{0.3^2}{0.5} = \frac{0.49}{0.5} + \frac{0.09}{0.5} = 0.98 + 0.18 = 1.16.$$

$$D_2(p\|q) = \frac{1}{2-1} \log_2(1.16) = \log_2(1.16) = 0.2141 \text{ bits.}$$

Rényi divergence of order $\alpha = 1/2$ (Bhattacharyya):

$$BC(p,q) = \sqrt{0.7 \times 0.5} + \sqrt{0.3 \times 0.5} = \sqrt{0.35} + \sqrt{0.15} = 0.5916 + 0.3873 = 0.9789.$$

$$D_{1/2}(p\|q) = \frac{1}{1/2-1} \log_2(0.9789) = -2 \log_2(0.9789) = -2(-0.0308) = 0.0615 \text{ bits.}$$

Comparison: $D_{1/2} = 0.0615 < D_{\mathrm{KL}}(\,\|\,) = 0.1184 < D_2 = 0.2141$, consistent with monotonicity in α. The ordering shows that the order-2 Rényi divergence is a strictly more conservative (larger) measure of the discrepancy between these two coin distributions than the KL divergence, while the order-$1/2$ value is more lenient.

Tsallis entropy and the q-deformed logarithm. A second, independently motivated generalization arises from statistical mechanics. Define the *q-logarithm* $\ln_q(x) = \frac{x^{1-q}-1}{1-q}$ for $x > 0$, $q \neq 1$, which reduces to the ordinary natural logarithm as $q \to 1$ (by the same L'Hôpital argument as before, applied to $x^{1-q} - 1$ over $1 - q$).

Tsallis Entropy

For $q \in \mathbb{R}$, $q \neq 1$, the *Tsallis entropy* (also called the q-entropy or the Havrda–Charvát entropy) of $X \sim p$ is

$$S_q(X) = \frac{1}{q-1}\left(1 - \sum_x p(x)^q\right) = -\sum_x p(x)^q \ln_q p(x) = \mathbb{E}_p[-\ln_q p(X)].$$

Relation to Rényi entropy. Both S_q and H_α (with $\alpha = q$) are monotone functions of the same quantity $Z_q := \sum_x p(x)^q$:

$$H_q(X) = \frac{1}{1-q} \log Z_q, \qquad S_q(X) = \frac{1}{q-1}(1 - Z_q).$$

Writing $H_q = \frac{1}{1-q}\log Z_q$ and $Z_q = 1-(q-1)S_q$, we obtain the exact correspondence

$$S_q(X) = \frac{1-e^{(1-q)H_q(X)}}{q-1}, \qquad H_q(X) = \frac{1}{1-q}\log(1-(q-1)S_q(X)).$$

The two families therefore carry exactly the same information — they are related by a strictly monotone reparametrization at each fixed q — but they behave very differently under composition of independent systems, which is the entire point of introducing Tsallis entropy separately.

Pseudo-additivity: the defining feature of Tsallis entropy. For Shannon entropy, independence gives additivity: $H(X,Y) = H(X)+H(Y)$ when $X \perp Y$. Rényi entropy also fails to be additive for $\alpha \neq 1$ in general (it is additive only because $Z_q(p\times r) = Z_q(p)Z_q(r)$ for product measures, which actually *does* hold — so $H_q(X,Y) = H_q(X)+H_q(Y)$ for independent X, Y too). Tsallis entropy, by contrast, satisfies the *pseudo-additivity* (or non-extensivity) relation

$$S_q(X,Y) = S_q(X) + S_q(Y) + (1-q)\,S_q(X)S_q(Y) \qquad (X \perp Y).$$

Proof. $Z_q(X,Y) = \sum_{x,y}(p(x)r(y))^q = Z_q(X)Z_q(Y)$ for independent X, Y. Substituting $Z_q(X) = 1-(q-1)S_q(X)$ and $Z_q(Y) = 1-(q-1)S_q(Y)$ into $Z_q(X,Y) = 1-(q-1)S_q(X,Y)$ gives

$$1-(q-1)S_q(X,Y) = (1-(q-1)S_q(X))(1-(q-1)S_q(Y)),$$

and expanding the right side and solving for $S_q(X,Y)$ yields the pseudo-additivity formula above.

For $q < 1$, the correction term $(1-q)S_q(X)S_q(Y) > 0$ is positive: the joint entropy of two independent systems *exceeds* the sum of their individual entropies — this is the *superextensive* regime, relevant to systems with long-range correlations where the whole genuinely contains more "surprise" than the sum of its parts would suggest. For $q > 1$ the correction is negative (*subextensive*). Only at $q = 1$ does ordinary additivity hold, which is precisely the Shannon case.

Statistical-mechanical motivation. Tsallis entropy arises as the entropy functional whose maximization under the constraints $\sum_x p(x) = 1$ and $\sum_x p(x)^q\,\varepsilon(x) = U$ (a generalized energy constraint) yields, via Lagrange multipliers, a power-law (Tsallis or q-exponential) distribution $p(x) \propto \left[1-(1-q)\beta\varepsilon(x)\right]_+^{1/(1-q)}$ rather than the ordinary Boltzmann–Gibbs exponential $p(x) \propto e^{-\beta\varepsilon(x)}$ obtained from maximizing Shannon entropy under a linear energy constraint. As $q \to 1$, the q-exponential $e_q(x) = [1+(1-q)x]_+^{1/(1-q)}$ reduces to e^x, and the Tsallis equilibrium reduces to the Boltzmann–Gibbs one. Power-law (heavy-tailed) equilibrium distributions are observed empirically in a range of complex systems — turbulence, certain plasmas, long-range gravitating systems, some financial return distributions — which motivates the use of S_q with $q \neq 1$ as the entropy functional for these systems' equilibrium statistics. We will revisit this q-exponential family in Volume III as an example of a non-regular exponential family with a non-trivial information geometry.

Example — Shannon, Rényi, and Tsallis entropy compared across q

Reuse $p = (0.50, 0.25, 0.15, 0.10)$, natural logarithms (nats) for this comparison.
Z_q values: $Z_{0.5} = \sum\sqrt{p(x)} = 1.9106$ (computed earlier in base-2 form; the value of $\sum p(x)^{0.5}$ is base-independent), $Z_1 = 1$ (by normalization), $Z_2 = \sum p(x)^2 = 0.345$.
Shannon entropy (nats): $H(X) = 1.743$ bits $\times \ln 2 = 1.208$ nats.

Tsallis entropy at $q = 0.5$:

$$S_{0.5}(X) = \frac{1 - Z_{0.5}}{0.5 - 1} = \frac{1 - 1.9106}{-0.5} = \frac{-0.9106}{-0.5} = 1.8212.$$

Tsallis entropy at $q = 2$:

$$S_2(X) = \frac{1 - Z_2}{2 - 1} = 1 - 0.345 = 0.655.$$

Tsallis entropy as $q \to 1$: by the L'Hôpital limit, $S_1(X) = H(X) = 1.208$ nats, sitting (as expected) between the $q = 0.5$ value 1.821 and the $q = 2$ value 0.655. Just as for Rényi entropy, *Tsallis entropy is monotonically decreasing in q* — but note that the *numerical values* $S_{0.5} = 1.821$ and $H_{0.5} = 1.868$ bits $\times \ln 2 / \log_2 e \approx 1.295$ nats are not equal (they cannot be, since S_q and H_q are related by a nonlinear reparametrization, not equality), even though both orderings agree qualitatively.

Min-entropy, guessing, and randomness extraction. We close this section by developing the operational meaning of H_∞ promised earlier.

Guessing Probability and Massey's Inequality

The *guessing probability* of $X \sim p$ is $P_{\text{guess}}(X) = \max_x p(x) = 2^{-H_\infty(X)}$: the probability that a single optimal guess (always guess the mode) is correct. Massey's guessing inequality relates the *expected number of guesses* $G(X)$ needed by an optimal sequential guessing strategy (guess outcomes in decreasing order of probability until correct) to Shannon entropy:

$$G(X) \geq \frac{1}{4} 2^{H(X)} + 1.$$

Interpretation. Min-entropy $H_\infty(X)$ governs the *single-shot* guessing probability exactly: $P_{\text{guess}}(X) = 2^{-H_\infty(X)}$ is an identity, not a bound. Shannon entropy, by contrast, governs only the *expected number of guesses* for the full sequential strategy, and only up to a constant-factor bound (Massey's inequality above; a matching upper bound $G(X) \leq 2^{H(X)} + 1$ also holds). This dichotomy — exact single-shot characterization by H_∞ versus average-case characterization by H — is the reason cryptographic security definitions (which must hold against an adversary who gets to make a *single* guess at a secret key, with no opportunity to retry) are phrased in terms of min-entropy rather than Shannon entropy: a source can have high Shannon entropy while still placing, say, probability 0.01 on a single easily-guessed value, which is catastrophic for a single-shot adversary but barely visible in the Shannon average.

Randomness extraction and the leftover hash lemma (preview). A randomness extractor is a function $\text{Ext} : \mathcal{X} \times \{0,1\}^d \to \{0,1\}^m$ that maps a high-min-entropy but non-uniform source X (together with a short uniform seed) to an output that is statistically close to uniform on $\{0,1\}^m$. The leftover hash lemma — proved using collision entropy H_2 via the Cauchy–Schwarz inequality applied to the collision probability of the hashed output — guarantees that a universal hash function extracts $m \approx H_2(X) - O(1)$ bits that are ε-close to uniform, for $\varepsilon = O(2^{-(H_2(X)-m)/2})$. Since $H_\infty(X) \leq H_2(X) \leq H(X)$, the extractable length is governed by the *worst-case* (smallest) of these quantities, reinforcing the theme that for adversarial / worst-case guarantees, the low-α... no — the *high*-α Rényi entropies (which are the smallest in the family) are the operationally relevant ones, while Shannon entropy ($\alpha = 1$) governs only average-case / asymptotic statements such as the source coding theo-

rem of Part 3.

Looking ahead. The Rényi divergence D_α introduced here will reappear in Volume III as the generator of the α-*connections* on a statistical manifold: the family of affine connections $\nabla^{(\alpha)}$ that interpolate between the exponential connection ($\alpha = 1$) and the mixture connection ($\alpha = -1$), with the Fisher information metric as the common Riemannian metric compatible with all of them. The non-extensivity of Tsallis entropy, meanwhile, foreshadows the discussion of *robust divergences* for heavy-tailed models, where power-divergences built from $\sum_x p(x)^\alpha q(x)^{1-\alpha}$-type integrals (rather than the logarithmic KL form) yield estimators with bounded influence functions — a theme developed at length in the companion volume on robustness and risk.

PART 2

Mutual Information and Divergence

KL divergence, mutual information, the data-processing inequality, sufficient statistics, and Pinsker's inequality.

2.1 Mutual Information

Mutual Information

The *mutual information* between discrete random variables X and Y is

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

Equivalently, $I(X;Y) = D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y))$.

Symmetry and non-negativity. $I(X;Y) = I(Y;X) \geq 0$, with equality iff $X \perp Y$. The symmetry follows directly from the definition; non-negativity follows from $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$.

Fundamental identity. The chain rule for mutual information:

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

These equalities yield the classical information diagram: $H(X)$ and $H(Y)$ overlap in $I(X;Y)$, with exclusive parts $H(X|Y)$ and $H(Y|X)$ and joint part $H(X,Y)$.

Conditional mutual information.

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = \sum_z p(z) I(X;Y|Z=z).$$

Chain rule for mutual information:

$$I(X_1, X_2;\, Y) = I(X_1;Y) + I(X_2;Y|X_1).$$

More generally, $I(X_1, \dots, X_n;\, Y) = \sum_{i=1}^{n} I(X_i;\, Y|X_1, \dots, X_{i-1})$.

Example — Mutual information in a binary channel

$X \in \{0,1\}$ uniform, $Y = X \oplus N$ where $N \sim$ Bernoulli(p) independent of X (binary symmetric channel with crossover probability p).

$$\begin{aligned} H(Y) &= H(X \oplus N) = h(p \star 1/2) = h(1/2) = 1 \text{ bits}, \\ H(Y|X) &= H(N|X) = H(N) = h(p), \\ I(X;Y) &= H(Y) - H(Y|X) = 1 - h(p). \end{aligned}$$

At $p = 0$: $I(X;Y) = 1$ bits (noiseless, Y determines X). At $p = 1/2$: $I(X;Y) = 0$ (Y is independent of X). At $p = 0.1$: $I(X;Y) = 1 - 0.469 = 0.531$ bits.

Mutual information and sufficient statistics. A statistic $T(X^n)$ is *sufficient* for parameter θ iff $I(\theta; X^n) = I(\theta; T(X^n))$: the statistic captures all the information about θ that the data contains.

Interaction information. For three variables X, Y, Z, the interaction information (also called co-information) is

$$I(X;Y;Z) = I(X;Y) - I(X;Y|Z),$$

which can be negative (indicating that conditioning on Z *increases* the mutual information between X and Y, a phenomenon called explaining away or Berkson's paradox).

2.2 KL Divergence and Its Properties

Kullback–Leibler Divergence

For probability distributions p and q on $\mathcal{X}$, the *KL divergence* (relative entropy) is

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

with the conventions $0 \log(0/q) = 0$ and $p \log(p/0) = +\infty$ for $p > 0$. For continuous densities: $D_{\mathrm{KL}}(f \,\|\, g) = \int f(x) \log(f(x)/g(x))\, dx$.

Non-negativity: Gibbs' inequality. $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$ for all p, q, with equality iff $p = q$ a.e.

Proof via Jensen's inequality.

$$-D_{\mathrm{KL}}(p \,\|\, q) = \sum_x p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_x p(x) \cdot \frac{q(x)}{p(x)} = \log \sum_x q(x) = \log 1 = 0,$$

using Jensen's inequality with the concave function log. Equality holds iff $q(x)/p(x)$ is constant, i.e., $p = q$.

Asymmetry. $D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$ in general. Example: $p = (0.9, 0.1)$, $q = (0.1, 0.9)$:

$$D_{\mathrm{KL}}(p \,\|\, q) = 0.9 \log \frac{0.9}{0.1} + 0.1 \log \frac{0.1}{0.9} = 0.9 \log 9 - 0.1 \log 9 = 0.8 \log 9 = 2.530 \text{ bits},$$
$$D_{\mathrm{KL}}(q \,\|\, p) = 0.1 \log 9 + 0.9 \log(1/9) = D_{\mathrm{KL}}(p \,\|\, q) \text{ (symmetric here by coincidence)}.$$

For asymmetric example: $p = (1/2, 1/4, 1/4)$, $q = (1/3, 1/3, 1/3)$:

$$D_{\mathrm{KL}}(p \,\|\, q) = \tfrac{1}{2} \log \tfrac{3}{2} + 2 \cdot \tfrac{1}{4} \log \tfrac{3}{4} = 0.292 - 0.415 = 0.0817 \text{ bits},$$
$$D_{\mathrm{KL}}(q \,\|\, p) = \tfrac{1}{3} \log \tfrac{2}{3} + 2 \cdot \tfrac{1}{3} \log \tfrac{4}{3} = -0.585/3 + 2 \cdot 0.415/3 = 0.0817 \text{ bits}.$$

Symmetric here again due to special structure; in general they differ.

KL divergence as a statistical test. By the Neyman–Pearson lemma, the log-likelihood ratio $\log(p(x)/q(x))$ is the optimal statistic for testing $H_0 : q$ vs $H_1 : p$. Under H_1 (true distribution p), the expected log-likelihood ratio per observation converges to $D_{\mathrm{KL}}(p \,\|\, q)$ by the law of large numbers. The KL divergence thus quantifies the statistical power of distinguishing p from q.

Stein's lemma. For n i.i.d. observations, the probability of incorrectly accepting H_0 under H_1 (type II error β_n) satisfies $\beta_n \doteq e^{-n D_{\mathrm{KL}}(p \,\|\, q)}$ as $n \to \infty$ (for any fixed type I error level $\alpha > 0$). This gives the operational meaning of KL divergence as an error exponent.

Example — KL divergence between two Gaussians

Let $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$.

$$D_{\mathrm{KL}}(p \,\|\, q) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

Derivation.

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) &= \mathbb{E}_p\left[\log \frac{p(X)}{q(X)}\right] = \mathbb{E}_p\left[\log \frac{\sigma_2}{\sigma_1} + \frac{(X-\mu_2)^2}{2\sigma_2^2} - \frac{(X-\mu_1)^2}{2\sigma_1^2}\right] \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\mathbb{E}_p[(X-\mu_2)^2]}{2\sigma_2^2} - \frac{\sigma_1^2}{2\sigma_1^2} \\ &= \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1-\mu_2)^2}{2\sigma_2^2} - \frac{1}{2}. \end{aligned}$$

Numerical instance: $\mu_1 = 0$, $\sigma_1 = 1$; $\mu_2 = 1$, $\sigma_2 = 2$:

$$D_{\mathrm{KL}}(p \,\|\, q) = \ln 2 + \frac{1+1}{8} - \frac{1}{2} = 0.693 + 0.25 - 0.5 = 0.443 \text{ nats} = 0.639 \text{ bits}.$$

Note the asymmetry: $D_{\mathrm{KL}}(q \,\|\, p) = \ln(1/2) + (4+1)/2 - 1/2 = -0.693 + 2 = 1.307$ nats $= 1.886$ bits.

f-divergences. KL divergence is a member of the f-divergence family:

$$D_f(p\|q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right),$$

where f is a convex function with $f(1) = 0$. KL: $f(t) = t \log t$. Total variation: $f(t) = |t-1|/2$. Hellinger: $f(t) = (\sqrt{t} - 1)^2$. All f-divergences satisfy $D_f(p\|q) \geq 0$ and the data-processing inequality.

2.3 The Data-Processing Inequality

Data-Processing Inequality (DPI)

If $X \to Y \to Z$ form a Markov chain (i.e., X and Z are conditionally independent given Y), then

$$I(X; Z) \leq I(X; Y).$$

More generally, for any stochastic function (deterministic or random) $Z = g(Y)$:

$$I(X; g(Y)) \leq I(X; Y).$$

Proof. By the chain rule for mutual information:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z) = I(X; Y) + I(X; Z|Y).$$

Since $X \to Y \to Z$ is a Markov chain, $I(X; Z|Y) = 0$ (knowing Y, Z gives no additional information about X). Therefore $I(X; Z) + I(X; Y|Z) = I(X; Y)$. Since $I(X; Y|Z) \geq 0$, we conclude $I(X; Z) \leq I(X; Y)$.

Interpretation. Processing data cannot create information. Any transformation of Y can only destroy or preserve the information that Y carries about X. This is the information-

theoretic basis for the impossibility of certain statistical tasks (e.g., a sufficient statistic can only preserve, never increase, mutual information).

Sufficient statistics. $T(X^n)$ is a sufficient statistic for θ iff $\theta \to X^n \to T(X^n) \to \theta$ forms a Markov chain, i.e., $I(\theta; T(X^n)) = I(\theta; X^n)$. By the DPI, $I(\theta; T) \le I(\theta; X^n)$ always; equality characterises sufficiency. The Fisher–Neyman factorization theorem provides the algebraic criterion.

Example — DPI in image compression

Original image X contains $n = 10^6$ pixels with entropy rate $\mathcal{H} = 8$ bits/pixel, so $H(X) = 8 \times 10^6$ bits. A compression algorithm g produces a $Y = g(X)$ of 500KB $= 4 \times 10^6$ bits. Query Z is a yes/no question about X (1 bit of information). By DPI: $I(X; Z) \le I(X; Y)$. But $I(X; Y) \le H(Y) \le 4 \times 10^6$ bits, so the compressed image retains at most 4×10^6 bits of information about the original. Any specific question about X that requires more than the compressed representation cannot be answered. This formalises the information loss in lossy compression.

Degraded channels. A channel $p(z|x)$ is *degraded* with respect to $p(y|x)$ if $X \to Y \to Z$ is a Markov chain. Then $C_{XZ} \le C_{XY}$: the degraded channel has lower capacity. This motivates the broadcast channel model in network information theory.

DPI for KL divergence. For any stochastic kernel T (Markov kernel) and distributions p, q:

$$D_{\mathrm{KL}}(Tp \,\|\, Tq) \le D_{\mathrm{KL}}(p \,\|\, q)\,,$$

where $(Tp)(y) = \sum_x T(y|x)p(x)$. This is the KL-divergence version of the DPI. **Proof.** By the log-sum inequality:

$$D_{\mathrm{KL}}(Tp \,\|\, Tq) = \sum_y (Tp)(y) \log \frac{(Tp)(y)}{(Tq)(y)} \le \sum_{x,y} p(x)T(y|x) \log \frac{p(x)T(y|x)}{q(x)T(y|x)} = D_{\mathrm{KL}}(p \,\|\, q)\,.$$

2.4 Pinsker's Inequality and Total Variation

Total Variation Distance

The *total variation* (TV) distance between distributions p and q on $\mathcal{X}$ is

$$\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \sum_{x \in \mathcal{X}} |p(x) - q(x)| = \max_{A \subseteq \mathcal{X}} |p(A) - q(A)|.$$

Pinsker's Inequality

$$\|p - q\|_{\mathrm{TV}}^2 \le \frac{\ln 2}{2} D_{\mathrm{KL}}(p \,\|\, q) = \frac{1}{2} D_{\mathrm{KL}}(p \,\|\, q) \text{ nats.}$$

In terms of bits: $\|p - q\|_{\mathrm{TV}}^2 \le \frac{\ln 2}{2} D_{\mathrm{KL}}(p \,\|\, q)_{\text{bits}}$.

Proof via a two-element reduction. Define $A^* = \{x : p(x) \ge q(x)\}$. Then $\|p - q\|_{\mathrm{TV}} = p(A^*) - q(A^*)$. Let $\alpha = p(A^*)$ and $\beta = q(A^*)$; define the two-point distributions $\tilde{p} = (\alpha, 1-\alpha)$ and $\tilde{q} = (\beta, 1-\beta)$ on $\{A^*, (A^*)^c\}$. By the data-processing inequality for KL:

$$D_{\mathrm{KL}}(\tilde{p} \,\|\, \tilde{q}) \le D_{\mathrm{KL}}(p \,\|\, q)\,.$$

For two-point distributions, Pinsker's inequality is proved directly: define $f(\alpha, \beta) = D_{\mathrm{KL}}(\tilde{p} \,\|\, \tilde{q}) - 2(\alpha - \beta)^2$ and show $f \geq 0$ by calculus. At $\alpha = \beta$, $f = 0$ and $\partial f / \partial \alpha = 0$; the second-order expansion of f in $(\alpha - \beta)$ shows $f \geq 0$. The general case follows.

Applications.

- **Lower bounds on KL divergence**: if two distributions are far in TV, they must be far in KL. This converts TV lower bounds (often easy to establish) into KL lower bounds (needed for coding theorems).
- **Hypothesis testing**: total variation directly measures the best possible error probability in distinguishing p from q. Pinsker links this to the KL-based error exponent.
- **MCMC convergence**: bounding $\|p_t - \pi\|_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}(p_t \,\|\, \pi)\,/2}$ via Pinsker converts KL bounds (often from Lyapunov arguments) into TV convergence rates.

Example — Pinsker's inequality for Bernoulli distributions

Let $p = \text{Bernoulli}(\alpha)$ and $q = \text{Bernoulli}(\beta)$ with $\alpha = 0.7$, $\beta = 0.5$.

$$\|p - q\|_{\mathrm{TV}} = |\alpha - \beta| = 0.2.$$

Pinsker lower bound: $\|p - q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{KL}}(p \,\|\, q)$ nats. $D_{\mathrm{KL}}(p \,\|\, q) = 0.7 \ln(0.7/0.5) + 0.3 \ln(0.3/0.5) = 0.7 \times 0.336 + 0.3 \times (-0.511) = 0.235 - 0.153 = 0.082$ nats. Pinsker bound: $0.04 \leq 0.041$. Tight! The Pinsker bound is sharp to constant factors for small perturbations.

Bretagnolle–Huber inequality (improvement on Pinsker).

$$\|p - q\|_{\mathrm{TV}} \leq \sqrt{1 - e^{-D_{\mathrm{KL}}(p \,\|\, q)}}.$$

This is tighter than Pinsker for large KL divergences and is increasingly used in minimax lower bound proofs.

Hellinger distance and its relation to KL. The Hellinger distance $H^2(p, q) = \frac{1}{2} \sum_x (\sqrt{p(x)} - \sqrt{q(x)})^2$ satisfies:

$$H^2(p, q) \leq \|p - q\|_{\mathrm{TV}} \leq \sqrt{2} H(p, q), \quad H^2(p, q) \leq \frac{1}{2} D_{\mathrm{KL}}(p \,\|\, q)\,.$$

The chain $H^2 \leq \|TV\| \leq \sqrt{KL/2} \cdot \sqrt{2}$ localises the divergences in a nested ordering.

2.5 f-Divergences and Variational Representations

KL divergence is the canonical measure of discrepancy between two distributions, but it is itself a member of a much larger family — the *f-divergences* — all of which share the two properties that make KL divergence useful: non-negativity with equality iff $p = q$, and the data-processing inequality. This section develops the f-divergence framework, places KL divergence, total variation, χ^2-divergence, and Hellinger distance within it as special cases, and then derives the variational (dual) representations of these divergences — a perspective that is the mathematical foundation of modern generative modeling (the f-GAN framework approximates an f-divergence between a model and the data distribution using only samples, via exactly the dual representation derived here) and that reappears in Volume III as the source of the dualistic geometric structure of statistical manifolds.

f-Divergence

Let $f : (0, \infty) \to \mathbb{R}$ be convex with $f(1) = 0$. For probability distributions p, q on $\mathcal{X}$ with $p \ll q$, the *f-divergence* of p from q is

$$D_f(p\|q) = \sum_{x \in \mathcal{X}} q(x)\, f\left(\frac{p(x)}{q(x)}\right) = \mathbb{E}_q\left[f\left(\frac{p(X)}{q(X)}\right)\right].$$

Non-negativity from Jensen's inequality. Since f is convex and $\mathbb{E}_q\left[\frac{p(X)}{q(X)}\right] = \sum_x q(x)\frac{p(x)}{q(x)} = \sum_x p(x) = 1$, Jensen's inequality gives

$$D_f(p\|q) = \mathbb{E}_q\left[f\left(\frac{p(X)}{q(X)}\right)\right] \geq f\left(\mathbb{E}_q\left[\frac{p(X)}{q(X)}\right]\right) = f(1) = 0.$$

Equality in Jensen's inequality holds iff $p(X)/q(X)$ is almost-surely constant under q, i.e. $p(x)/q(x) = 1$ for all x with $q(x) > 0$, i.e. $p = q$ (given $p \ll q$). If f is *strictly* convex at 1, equality $D_f(p\|q) = 0$ holds if and only if $p = q$. This single computation — already familiar from the proof that $D_{\mathrm{KL}}(p \,\|\, q) \geq 0$ in Section 2.2 — is the source of non-negativity for *every* member of the f-divergence family simultaneously: the choice of f only affects the *shape* of the divergence (how it weights different likelihood-ratio regimes), never the basic non-negativity property, which is inherited wholesale from the convexity of f.

The classical examples as f-divergences. Each familiar divergence corresponds to a specific generator f:

KL divergence. Take $f(t) = t \log t$ (convex, since $f''(t) = 1/t > 0$, and $f(1) = 0$). Then

$$D_f(p\|q) = \sum_x q(x) \cdot \frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} = \sum_x p(x) \log \frac{p(x)}{q(x)} = D_{\mathrm{KL}}(p \,\|\, q)\,.$$

Reverse KL divergence. Take $f(t) = -\log t$ (convex, $f(1) = 0$). Then $D_f(p\|q) = \sum_x q(x)\left(-\log \frac{p(x)}{q(x)}\right) = \sum_x q(x) \log \frac{q(x)}{p(x)} = D_{\mathrm{KL}}(q \,\|\, p)$. This shows that $D_{\mathrm{KL}}(p \,\|\, q)$ and $D_{\mathrm{KL}}(q \,\|\, p)$ are *both* f-divergences, but for different generators — they are not related by simply swapping arguments in a single fixed D_f, which is why KL divergence is asymmetric while the f-divergence *framework* treats p and q on an equal conceptual footing (the asymmetry is entirely encoded in the choice of f).

Total variation. Take $f(t) = \frac{1}{2}|t - 1|$ (convex — it is a scaled absolute value — and $f(1) = 0$). Then

$$D_f(p\|q) = \sum_x q(x) \cdot \frac{1}{2}\left|\frac{p(x)}{q(x)} - 1\right| = \frac{1}{2}\sum_x |p(x) - q(x)| = \|p - q\|_{\mathrm{TV}}.$$

χ^2-divergence. Take $f(t) = (t-1)^2$ (convex, $f(1) = 0$, $f''(t) = 2 > 0$). Then

$$D_f(p\|q) = \sum_x q(x)\left(\frac{p(x)}{q(x)} - 1\right)^2 = \sum_x \frac{(p(x) - q(x))^2}{q(x)} =: \chi^2(p, q).$$

This is the *Pearson chi-squared divergence*, familiar from the chi-squared goodness-of-fit test, now revealed to be an f-divergence.

Squared Hellinger distance. Take $f(t) = (\sqrt{t} - 1)^2$ (convex on $(0, \infty)$, $f(1) = 0$). Then

$$D_f(p\|q) = \sum_x q(x)\left(\sqrt{\frac{p(x)}{q(x)}} - 1\right)^2 = \sum_x \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 = 2H^2(p, q),$$

recovering (up to the factor of 2 convention used in Section 2.4) the squared Hellinger distance encountered earlier.

Jensen–Shannon divergence. Take $f(t) = t\log\frac{2t}{t+1} + \log\frac{2}{t+1}$ (convex, $f(1)=0$, and bounded — $D_f \le \log 2$ always). This f-divergence is symmetric in p, q despite being built from the asymmetric KL generator, because it is in fact $\mathrm{JS}(p,q) = \frac{1}{2}D_{\mathrm{KL}}(p\,\|\,m) + \frac{1}{2}D_{\mathrm{KL}}(q\,\|\,m)$ where $m = \frac{1}{2}(p+q)$; verifying the equivalence of the two formulas is a short exercise in algebra left to the reader as a check of the definitions above.

Example — Five f-divergences between two Bernoulli distributions

Let $p = \mathrm{Bernoulli}(0.7)$, $q = \mathrm{Bernoulli}(0.5)$ as in the previous section, base-2 logarithms where relevant.
Total variation: $\|p-q\|_{\mathrm{TV}} = \frac{1}{2}(|0.7-0.5| + |0.3-0.5|) = \frac{1}{2}(0.2+0.2) = 0.2$.
KL divergence: from the previous section, $D_{\mathrm{KL}}(p\,\|\,q) = 0.1184$ bits.
Reverse KL: $D_{\mathrm{KL}}(q\,\|\,p) = 0.5\log_2\frac{0.5}{0.7} + 0.5\log_2\frac{0.5}{0.3} = 0.5(-0.485) + 0.5(0.737) = -0.2425 + 0.3685 = 0.1260$ bits. Note $D_{\mathrm{KL}}(p\,\|\,q) \neq D_{\mathrm{KL}}(q\,\|\,p)$ (0.1184 vs 0.1260) — the asymmetry of KL divergence, here made concrete: it costs slightly more (in expected log-likelihood-ratio terms) to mistake q for p when the truth is q than the reverse.
χ^2-divergence:

$$\chi^2(p,q) = \frac{(0.7-0.5)^2}{0.5} + \frac{(0.3-0.5)^2}{0.5} = \frac{0.04}{0.5} + \frac{0.04}{0.5} = 0.08 + 0.08 = 0.16.$$

Squared Hellinger: $H^2(p,q) = \frac{1}{2}\left[(\sqrt{0.7}-\sqrt{0.5})^2 + (\sqrt{0.3}-\sqrt{0.5})^2\right] = \frac{1}{2}\left[(0.8367-0.7071)^2 + (0.5477-0.7071)^2\right] = \frac{1}{2}[0.01679 + 0.02541] = \frac{1}{2}(0.0422) = 0.0211$.
Cross-checking the chain of inequalities. The chain $H^2 \le \mathrm{TV} \le \sqrt{2H^2}$ from Section 2.4 gives $0.0211 \le 0.2 \le \sqrt{0.0422} = 0.2054$: both inequalities hold, with the upper one nearly tight. Pinsker's inequality $\mathrm{TV} \le \sqrt{D_{\mathrm{KL}}(\,\|\,)/2 \cdot \ln 2}$ (converting the bits-based KL to nats: $D_{\mathrm{KL}}(\,\|\,) = 0.1184 \times \ln 2 = 0.0821$ nats) gives $\mathrm{TV} \le \sqrt{0.0821/2} = \sqrt{0.04105} = 0.2026$, again consistent with the actual value 0.2.

Joint convexity of f-divergences. A key structural property: $(p,q) \mapsto D_f(p\|q)$ is *jointly convex* in the pair (p,q) — convex as a function on the product of probability simplices, not merely separately convex in each argument. This follows from the joint convexity of the *perspective function* $(u,v) \mapsto vf(u/v)$ for $v>0$, which is a standard fact: if f is convex, its perspective $g(u,v) = vf(u/v)$ is jointly convex on $\{v>0\}$. (Proof: g is the function whose epigraph is the conic hull of the epigraph of f, and conic hulls of convex sets are convex.) Summing the jointly convex functions $q(x)f(p(x)/q(x))$ over x preserves joint convexity. Joint convexity is the property used in Section 2.3 to prove the convexity of KL divergence in (p,q), and it generalizes verbatim to every f-divergence.

Data-processing inequality, revisited in full generality. For *any* convex f with $f(1)=0$, and any Markov kernel (channel) K applied to produce $p_Y = Kp_X$, $q_Y = Kq_X$ from p_X, q_X,

$$D_f(p_Y\|q_Y) \le D_f(p_X\|q_X).$$

Proof sketch. Write $D_f(p_Y\|q_Y) = \sum_y q_Y(y) f\left(\frac{p_Y(y)}{q_Y(y)}\right)$. Since $p_Y(y) = \sum_x K(y|x)p_X(x)$ and $q_Y(y) = \sum_x K(y|x)q_X(x)$, the ratio $\frac{p_Y(y)}{q_Y(y)}$ is a $q_X(\cdot|y)$-weighted average (via Bayes' rule on the channel, i.e. a *mixture*) of the ratios $\frac{p_X(x)}{q_X(x)}$ over x contributing to y. Applying Jensen's inequality to f applied to this average, then summing over y with weights $q_Y(y)$, and using Fubini to swap the order of summation, recovers $\sum_x q_X(x) f(p_X(x)/q_X(x)) = D_f(p_X\|q_X)$ as

an upper bound. The details mirror exactly the proof for KL divergence in Section 2.3, with $t \log t$ replaced by the general convex f.

This single inequality is the reason that *every* f-divergence — total variation, χ^2, Hellinger, Jensen–Shannon, and the Rényi divergences of Section 1.6 (which, while not literally of the form D_f for a single fixed convex f independent of α, are monotone transformations of f-divergences for the family $f_\alpha(t) = \frac{t^\alpha - 1}{\alpha - 1}$) — shares the qualitative behavior that processing cannot increase distinguishability. This is the information-theoretic content behind the maxim "post-processing cannot create information."

Variational representation: the convex conjugate. The second major theme of this section is that every f-divergence admits a *dual* representation as a supremum over functions, rather than a direct sum/integral over $\mathcal{X}$. This is the route by which f-divergences become estimable from samples alone (without knowing p or q explicitly), which is essential when p is, say, the distribution induced by a deep generative model.

Convex Conjugate (Legendre–Fenchel Transform)

For a convex function $f : (0, \infty) \to \mathbb{R}$, its *convex conjugate* is

$$f^*(s) = \sup_{t>0} \{st - f(t)\}, \qquad s \in \operatorname{dom}(f^*).$$

Variational (Dual) Representation of f-Divergences

Since $f(t) = \sup_s \{st - f^*(s)\}$ (biconjugacy, when f is closed convex — which holds for all the generators above), we have, for any x,

$$f\left(\frac{p(x)}{q(x)}\right) = \sup_{s \in \operatorname{dom}(f^*)} \left\{ s \cdot \frac{p(x)}{q(x)} - f^*(s) \right\}.$$

Choosing a function $T : \mathcal{X} \to \operatorname{dom}(f^*)$ (rather than a single scalar s, since the optimal s may depend on x) and substituting,

$$D_f(p\|q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) \geq \sum_x q(x) \left[T(x)\frac{p(x)}{q(x)} - f^*(T(x)) \right] = \sum_x p(x)T(x) - \sum_x q(x) f^*(T(x)),$$

i.e.

$$D_f(p\|q) \geq \mathbb{E}_p[T(X)] - \mathbb{E}_q[f^*(T(X))] \qquad \text{for every } T : \mathcal{X} \to \operatorname{dom}(f^*),$$

with equality when $T(x) = f'(p(x)/q(x))$ pointwise (the maximizer of the pointwise Fenchel identity). Taking the supremum over T,

$$D_f(p\|q) = \sup_{T : \mathcal{X} \to \operatorname{dom}(f^*)} \Big(\mathbb{E}_p[T(X)] - \mathbb{E}_q[f^*(T(X))] \Big).$$

Specialization to KL divergence: the Donsker–Varadhan formula. For $f(t) = t \log t$, the conjugate is $f^*(s) = \sup_{t>0}\{st - t\log t\}$. Differentiating the inner expression in t: $\frac{\partial}{\partial t}(st - t\log t) = s - \log t - 1 = 0 \Rightarrow t = e^{s-1}$, giving $f^*(s) = se^{s-1} - e^{s-1}\log(e^{s-1}) = se^{s-1} - e^{s-1}(s-1) = e^{s-1}$. So $f^*(s) = e^{s-1}$, and the variational formula becomes

$$D_{\mathrm{KL}}(p \,\|\, q) = \sup_{T : \mathcal{X} \to \mathbb{R}} \Big(\mathbb{E}_p[T(X)] - \mathbb{E}_q[e^{T(X)-1}] \Big).$$

A standard reparametrization $T(x) = 1 + \log g(x)$ for $g > 0$ turns this into the more familiar

Donsker–Varadhan representation

$$D_{\mathrm{KL}}(p\,\|\,q) = \sup_{g:\mathcal{X}\to(0,\infty)} \Big(\mathbb{E}_p[\log g(X)] - \log \mathbb{E}_q[g(X)]\Big),$$

which is verified by substituting $T = 1 + \log g$ into the previous display and simplifying $\mathbb{E}_q[e^{T-1}] = \mathbb{E}_q[g]$, then noting that $\sup_c(\mathbb{E}_p[\log g] + c - \mathbb{E}_q[g]e^{c-1})$ over the remaining additive constant c is maximized at $c = 1 - \log \mathbb{E}_q[g]$, giving the $-\log \mathbb{E}_q[g(X)]$ term. The Donsker–Varadhan formula expresses $D_{\mathrm{KL}}(p\,\|\,q)$ as a supremum over arbitrary positive functions g of a quantity computable using only samples from p and q (empirical averages of $\log g(X)$ and $g(X)$) — this is precisely the principle behind mutual-information neural estimation (MINE) and related sample-based divergence estimators, where g (or T) is parametrized by a neural network and the supremum is approximated by gradient ascent.

Example — Variational lower bound for χ^2-divergence

For $f(t) = (t-1)^2$, compute $f^*(s) = \sup_{t>0}\{st-(t-1)^2\}$. Differentiating: $s-2(t-1) = 0 \Rightarrow t = 1 + s/2$, giving $f^*(s) = s(1+s/2) - (s/2)^2 = s + s^2/2 - s^2/4 = s + s^2/4$. So the variational formula reads

$$\chi^2(p,q) = \sup_{T:\mathcal{X}\to\mathbb{R}} \left(\mathbb{E}_p[T(X)] - \mathbb{E}_q[T(X)] - \tfrac{1}{4}\mathbb{E}_q[T(X)^2]\right).$$

Sanity check with $p = \text{Bernoulli}(0.7)$, $q = \text{Bernoulli}(0.5)$, $T(1) = a$, $T(0) = b$. We know the true value $\chi^2(p,q) = 0.16$ from the previous example. Take the *optimal* witness, which by the equality condition $T(x) = f'(p(x)/q(x)) = 2(p(x)/q(x) - 1)$: $T(1) = 2(0.7/0.5 - 1) = 2(0.4) = 0.8$, $T(0) = 2(0.3/0.5 - 1) = 2(-0.4) = -0.8$. Then

$$\begin{aligned}
\mathbb{E}_p[T] &= 0.7(0.8) + 0.3(-0.8) = 0.56 - 0.24 = 0.32,\\
\mathbb{E}_q[T] &= 0.5(0.8) + 0.5(-0.8) = 0,\\
\mathbb{E}_q[T^2] &= 0.5(0.64) + 0.5(0.64) = 0.64.
\end{aligned}$$

$$\mathbb{E}_p[T] - \mathbb{E}_q[T] - \tfrac{1}{4}\mathbb{E}_q[T^2] = 0.32 - 0 - \tfrac{1}{4}(0.64) = 0.32 - 0.16 = 0.16,$$

exactly recovering $\chi^2(p,q) = 0.16$, confirming that this T achieves the supremum (as the theory predicts) and illustrating that any *other* choice of T would yield a value ≤ 0.16 — a genuine lower bound usable when p, q are unknown but samplable.

The f-GAN connection. In generative adversarial training, a generator produces samples from a model distribution p_θ, and the goal is to make p_θ close to the data distribution q_{data} in some divergence. The f-GAN framework (Nowozin, Cseke, Tomioka, 2016) trains a "discriminator" network T_ω to approximate the supremum in $D_f(q_{\text{data}}\|p_\theta) \approx \sup_\omega(\mathbb{E}_{q_{\text{data}}}[T_\omega] - \mathbb{E}_{p_\theta}[f^*(T_\omega)])$, while the generator's parameters θ are updated to *minimize* this same quantity — exactly the minimax structure $\min_\theta \max_\omega(\cdots)$ familiar from the original GAN formulation, which corresponds to the choice $f(t) = t\log t - (t+1)\log(t+1)$ (up to additive constants), an f-divergence closely related to but distinct from the Jensen–Shannon divergence. Different choices of f — χ^2, reverse KL, Hellinger — yield different GAN variants with different sample-quality/diversity trade-offs, all unified by the single variational formula derived above. This is one of the most direct lines from the abstract convex-analytic machinery of this section to a widely used machine-learning algorithm.

Looking ahead: duality and information geometry. The Legendre–Fenchel duality $f \leftrightarrow f^*$ that produced the variational formulas above is the same duality that, in Volume III, will relate the two natural coordinate systems on a statistical manifold — the *expectation parameters*

(means of sufficient statistics) and the *natural parameters* (exponential-family canonical parameters) — via the Legendre transform of the log-partition function (cumulant generating function), which is itself a convex function of Fenchel-conjugate pairs in exactly the sense developed here. The f-divergence framework is thus not merely a taxonomy of distance-like quantities, but the shadow, on the space of distributions, of a convex-duality structure that becomes the geometric backbone of the exponential-family manifold.

2.6 Maximum Entropy, Exponential Families, and Lagrange Duality

We have now met entropy as an axiomatically characterized measure of uncertainty (Part 1), divergence as a measure of distance between distributions arising from a general convex-analytic framework (Section 2.5), and the empirical-distribution machinery that connects probability of deviation to KL divergence (Section 3.6, encountered out of the linear reading order but logically adjacent). This section ties these threads together via a single variational principle — *maximum entropy* — that simultaneously (a) explains why the familiar parametric families (Gaussian, exponential, Poisson, geometric, ...) are the "natural" choices in statistical modeling, (b) reveals entropy and the log-partition function as Legendre-conjugate convex functions, and (c) sets up, in concrete computational terms, the dually-flat geometric structure that Volume III develops as the geometry of statistical manifolds.

The Maximum Entropy Principle

Given a collection of functions ("features" or "sufficient statistics") $T_1, \dots, T_k : \mathcal{X} \to \mathbb{R}$ and target values $b_1, \dots, b_k \in \mathbb{R}$, the *maximum entropy distribution* subject to the moment constraints $\mathbb{E}_p[T_j(X)] = b_j$ for $j = 1, \dots, k$ is

$$p^* = \arg\max_{p \in \Delta(\mathcal{X})} H(p) \quad \text{subject to} \quad \sum_x p(x) T_j(x) = b_j \ (j = 1, \dots, k), \quad \sum_x p(x) = 1.$$

Motivation: the least-informative distribution consistent with what is known. If all we know about a random variable X is a finite set of moment values $b_1, \dots, b_k$ (say, its mean, or its mean and variance), then among all distributions consistent with those moments, the maximum-entropy one is — by the operational interpretation of entropy from Part 1 — the one that is hardest to compress, i.e. the one that assumes the *least* additional structure beyond the stated constraints. This is Jaynes' principle of maximum entropy (1957): it is not a physical law but a principle of inference, selecting the distribution that is maximally noncommittal about everything not specified by the constraints.

Solving the constrained optimization via Lagrange multipliers. Form the Lagrangian

$$\mathcal{L}(p, \boldsymbol{\lambda}, \nu) = -\sum_x p(x) \log p(x) - \sum_{j=1}^{k} \lambda_j \left(\sum_x p(x) T_j(x) - b_j \right) - \nu \left(\sum_x p(x) - 1 \right),$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_k)$ are multipliers for the moment constraints and ν for normalization. Differentiate with respect to $p(x)$ for each fixed x and set to zero:

$$\frac{\partial \mathcal{L}}{\partial p(x)} = -\log p(x) - 1 - \sum_{j=1}^{k} \lambda_j T_j(x) - \nu = 0 \implies \log p(x) = -1 - \nu - \sum_{j=1}^{k} \lambda_j T_j(x).$$

Therefore

$$p^*(x) = \exp\Big(-1 - \nu - \textstyle\sum_j \lambda_j T_j(x)\Big) = \frac{1}{Z(\boldsymbol{\lambda})} \exp\Big(-\textstyle\sum_j \lambda_j T_j(x)\Big),$$

where $Z(\boldsymbol{\lambda}) = \sum_x e^{-\sum_j \lambda_j T_j(x)}$ absorbs the constant $e^{1+\nu}$ and enforces normalization. *The maximum-entropy distribution subject to moment constraints is always a member of an exponential family* — this single computation is the origin of every "natural" parametric family encountered in statistics.

Exponential Family

A family of distributions $\{p_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta\subseteq\mathbb{R}^k}$ on $\mathcal{X}$ is an *exponential family* with *natural parameters* $\boldsymbol{\theta}$, *sufficient statistics* $T(x) = (T_1(x), \dots, T_k(x))$, and *log-partition function* (or cumulant generating function) $A(\boldsymbol{\theta})$ if

$$p_{\boldsymbol{\theta}}(x) = h(x)\exp(\boldsymbol{\theta}^\top T(x) - A(\boldsymbol{\theta})), \qquad A(\boldsymbol{\theta}) = \log \sum_x h(x) e^{\boldsymbol{\theta}^\top T(x)}$$

($h(x)$ a fixed base measure, often $h \equiv 1$ in the discrete case considered here; the sign convention $\boldsymbol{\theta} = -\boldsymbol{\lambda}$ relative to the Lagrangian derivation above is the standard one and absorbed into A).

Convexity of the log-partition function. $A(\boldsymbol{\theta})$ is convex (in fact, strictly convex on its domain unless the T_j are affinely dependent) — this is the single most important analytic property of exponential families, and it follows from Hölder's inequality exactly as the convexity of $f(\alpha) = \log \sum_x p(x)^\alpha$ did in Section 1.6. To see this directly, compute the Hessian: writing $Z(\boldsymbol{\theta}) = \sum_x h(x) e^{\boldsymbol{\theta}^\top T(x)} = e^{A(\boldsymbol{\theta})}$,

$$\frac{\partial A}{\partial \theta_i} = \frac{1}{Z}\sum_x h(x) T_i(x) e^{\boldsymbol{\theta}^\top T(x)} = \mathbb{E}_{\boldsymbol{\theta}}[T_i(X)],$$

$$\frac{\partial^2 A}{\partial \theta_i \partial \theta_j} = \mathbb{E}_{\boldsymbol{\theta}}[T_i T_j] - \mathbb{E}_{\boldsymbol{\theta}}[T_i]\mathbb{E}_{\boldsymbol{\theta}}[T_j] = \operatorname{Cov}_{\boldsymbol{\theta}}(T_i(X), T_j(X)).$$

The Hessian of A is the covariance matrix of the sufficient statistics under $p_{\boldsymbol{\theta}}$, which is positive semi-definite (and positive definite unless some linear combination of the T_j is almost surely constant). Hence A is convex. This single fact — *the log-partition function's gradient gives the mean parameters, and its Hessian gives their covariance* — is the computational engine behind maximum-likelihood estimation in exponential families (the log-likelihood is concave in $\boldsymbol{\theta}$, so gradient ascent / Newton's method converges to the unique maximum) and is the source of the Fisher information matrix that will become the Riemannian metric of Volume III: $g_{ij}(\boldsymbol{\theta}) = \partial^2 A/\partial\theta_i\partial\theta_j$ *is* the Fisher information metric in natural coordinates, a fact we derive independently (via the score function) in Volume III but which is already visible here as a direct consequence of differentiating the normalizing constant.

Example — The Gaussian as a maximum-entropy distribution

Setup. On $\mathcal{X} = \mathbb{R}$ (continuous case — the sums above become integrals, and entropy becomes differential entropy $h(p)$ from Section 1.4), impose the two constraints $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \mu^2+\sigma^2$ (equivalently, $\operatorname{Var}(X) = \sigma^2$). The sufficient statistics are $T_1(x) = x$, $T_2(x) = x^2$.

Solving. The maxent density has the form $p^*(x) \propto \exp(\lambda_1 x + \lambda_2 x^2)$. For this to be normalizable, $\lambda_2 < 0$; write $\lambda_2 = -\frac{1}{2\sigma^2}$ and complete the square: $\lambda_1 x + \lambda_2 x^2 = -\frac{1}{2\sigma^2}(x - \sigma^2\lambda_1)^2 + \frac{\sigma^2\lambda_1^2}{2}$. Matching $\sigma^2\lambda_1 = \mu$ gives $\lambda_1 = \mu/\sigma^2$, and the density becomes $p^*(x) \propto \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ — the Gaussian $\mathcal{N}(\mu, \sigma^2)$.

Verification via entropy comparison. From Section 1.4, $h(\mathcal{N}(\mu,\sigma^2)) = \frac{1}{2}\log(2\pi e\sigma^2)$. Any other density q with the same mean and variance satisfies $h(q) \le h(\mathcal{N}(\mu,\sigma^2))$

— this is exactly the maximum-entropy property, and it was proved directly (via $D_{\mathrm{KL}}(q \,\|\, \mathcal{N}) \geq 0$) in Section 1.4's discussion of the maximum entropy theorem for continuous distributions. The Lagrangian derivation here *explains* why that particular KL-divergence computation worked: $\mathcal{N}(\mu, \sigma^2)$ is, by construction, the unique exponential-family distribution with sufficient statistics (x, x^2) matching the given moments, and the general theory guarantees it maximizes entropy among *all* distributions (not just other exponential-family members) with those moments.

Example — The exponential and geometric distributions as maximum entropy

Exponential distribution. On $\mathcal{X} = [0, \infty)$ with the single constraint $\mathbb{E}[X] = 1/\lambda$ (mean fixed), $T_1(x) = x$, the maxent density is $p^*(x) \propto e^{\theta x}$ with $\theta < 0$ for normalizability; writing $\theta = -\lambda$, $p^*(x) = \lambda e^{-\lambda x}$ — the exponential distribution $\mathrm{Exp}(\lambda)$. This is the continuous analogue of the geometric distribution below, and is the maxent distribution on $[0, \infty)$ given only a mean constraint — explaining its ubiquity as a model for waiting times and lifetimes when only the mean lifetime is known.
Geometric distribution. On $\mathcal{X} = \{0, 1, 2, ...\}$ with $\mathbb{E}[X] = m$ fixed, $T_1(x) = x$, the maxent distribution is $p^*(x) \propto e^{\theta x} = r^x$ for $r = e^\theta \in (0, 1)$ — the geometric distribution $p^*(x) = (1 - r)r^x$ with mean $\mathbb{E}[X] = r/(1 - r) = m$, giving $r = m/(m + 1)$. **Numerical check, $m = 2$:** $r = 2/3$, so $p^*(x) = (1/3)(2/3)^x$. Entropy: $H(p^*) = -\sum_x p^*(x) \log_2 p^*(x)$. Using the general formula for the geometric distribution's entropy, $H = -\log_2(1-r) - \frac{r}{1-r} \log_2 r = -\log_2(1/3) - 2\log_2(2/3) = 1.585 - 2(-0.585) = 1.585 + 1.170 = 2.755$ bits. Any other distribution on $\{0, 1, 2, ...\}$ with mean 2 has entropy at most 2.755 bits — e.g. a Poisson(2) distribution, which has the same mean but is more concentrated (lighter tails) than the geometric, has strictly smaller entropy ($H(\mathrm{Poisson}(2)) \approx 2.20$ bits), consistent with the maxent property: among all mean-2 distributions on the non-negative integers, the geometric is the most spread out / least predictable.

KL divergence within an exponential family: the Bregman divergence of A. For two members $p_{\boldsymbol{\theta}_1}, p_{\boldsymbol{\theta}_2}$ of the same exponential family,

$$D_{\mathrm{KL}}\big(p_{\boldsymbol{\theta}_1} \,\|\, p_{\boldsymbol{\theta}_2}\big) = \mathbb{E}_{\boldsymbol{\theta}_1}\left[\log \frac{p_{\boldsymbol{\theta}_1}(X)}{p_{\boldsymbol{\theta}_2}(X)}\right] = \mathbb{E}_{\boldsymbol{\theta}_1}\left[(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top T(X) - A(\boldsymbol{\theta}_1) + A(\boldsymbol{\theta}_2)\right].$$

Using $\mathbb{E}_{\boldsymbol{\theta}_1}[T(X)] = \nabla A(\boldsymbol{\theta}_1) =: \boldsymbol{\mu}_1$ (the mean parameters corresponding to $\boldsymbol{\theta}_1$),

$$D_{\mathrm{KL}}\big(p_{\boldsymbol{\theta}_1} \,\|\, p_{\boldsymbol{\theta}_2}\big) = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \boldsymbol{\mu}_1 - A(\boldsymbol{\theta}_1) + A(\boldsymbol{\theta}_2) = A(\boldsymbol{\theta}_2) - A(\boldsymbol{\theta}_1) - \nabla A(\boldsymbol{\theta}_1)^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1).$$

The right-hand side is exactly $B_A(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$, the *Bregman divergence* of the convex function A between $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_1$ (compare the definition in Appendix B): *KL divergence between two exponential-family members equals the Bregman divergence of the log-partition function between their natural parameters*, with the arguments in the order matching the asymmetry of KL divergence ($D_{\mathrm{KL}}\big(p_{\theta_1} \,\|\, p_{\theta_2}\big)$ corresponds to $B_A(\theta_2, \theta_1)$ — note the swap). This identity converts an integral/sum over $\mathcal{X}$ (the definition of KL divergence) into a purely finite-dimensional computation involving A and its gradient — an enormous simplification whenever A has closed form.

Example — KL divergence between two Gaussians via the Bregman identity

For $\mathcal{N}(\mu, \sigma^2)$ written in natural-parameter form with $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/(2\sigma^2)$ (so $T(x) = (x, x^2)$), the log-partition function is $A(\boldsymbol{\theta}) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2) + \text{const}$. Rather than working through this two-dimensional Bregman computation in full generality, specialize to *fixed, equal variance* σ^2 and varying means μ_1, μ_2 — a one-dimensional exponential family in $\theta = \mu/\sigma^2$ with $A(\theta) = \sigma^2\theta^2/2$ (up to additive constants independent of θ, dropped). Then $\nabla A(\theta) = \sigma^2\theta = \mu$, and the Bregman divergence is

$$B_A(\theta_2, \theta_1) = A(\theta_2) - A(\theta_1) - A'(\theta_1)(\theta_2 - \theta_1) = \frac{\sigma^2\theta_2^2}{2} - \frac{\sigma^2\theta_1^2}{2} - \sigma^2\theta_1(\theta_2 - \theta_1) = \frac{\sigma^2}{2}(\theta_2 - \theta_1)^2 = \frac{(\mu_2 - \mu_1)^2}{2\sigma^2}.$$

Numerical check, $\mu_1 = 0, \mu_2 = 1, \sigma^2 = 1$**:** $B_A = \frac{1}{2}$. Direct computation of $D_{\mathrm{KL}}(\mathcal{N}(0,1) \,\|\, \mathcal{N}(1,1))$ via the general Gaussian KL formula $\frac{1}{2}\left[\frac{(\mu_1-\mu_2)^2}{\sigma^2} + \frac{\sigma^2}{\sigma^2} - 1 - \log 1\right] = \frac{1}{2}[1 + 1 - 1] = \frac{1}{2}$, matching exactly. The Bregman-divergence route required only evaluating a quadratic at two points and its derivative at one — no integration — illustrating the computational power of the exponential-family/Bregman correspondence for any pair of distributions within the same family.

The dual (mean) parametrization and Legendre duality. The map $\boldsymbol{\theta} \mapsto \boldsymbol{\mu} = \nabla A(\boldsymbol{\theta})$ is, by the strict convexity of A, a diffeomorphism from the natural-parameter space onto the *mean-parameter space* $\mathcal{M} = \{\mathbb{E}_p[T(X)] : p \in \Delta(\mathcal{X})\}$ (the convex hull of $\{T(x) : x \in \mathcal{X}\}$). Its inverse is $\boldsymbol{\mu} \mapsto \boldsymbol{\theta} = \nabla A^*(\boldsymbol{\mu})$, where A^* is the convex conjugate of A introduced in Section 2.5's general Fenchel-duality discussion, specialized here to the log-partition function. As recorded in Appendix B, $A^*(\boldsymbol{\mu}) = -H(p_{\boldsymbol{\mu}}) + \text{const}$ (up to sign and additive-constant conventions depending on whether $h(x) \equiv 1$): *the conjugate of the log-partition function is* (*minus*) *the entropy, expressed in mean-parameter coordinates*. This is the same Legendre pairing that appeared in Section 2.5 for general f-divergences, now specialized to the pair $(A, -H)$ — the cumulant generating function and the entropy are Legendre-conjugate convex functions on dual coordinate systems $(\boldsymbol{\theta}, \boldsymbol{\mu})$.

Two flat coordinate systems, one manifold. We thus have, for any exponential family, *two* natural coordinate systems — $\boldsymbol{\theta}$ (natural/canonical parameters) and $\boldsymbol{\mu}$ (mean/expectation parameters) — related by the gradient maps of mutually conjugate convex functions A and $A^* = -H$. Each coordinate system is "flat" with respect to a different affine connection (the *exponential connection*, for which $\boldsymbol{\theta}$-coordinates are affine, and the *mixture connection*, for which $\boldsymbol{\mu}$-coordinates are affine), and the KL divergence — expressible as a Bregman divergence in *either* coordinate system, $B_A(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1) = B_{A^*}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)$ — is the canonical "distance" compatible with both flat structures simultaneously. This is, in condensed form, the entire content of Volume III's opening chapters: the abstract apparatus of *dually flat spaces* is nothing more than the elementary Legendre duality between A and H derived here, promoted from a pair of convex functions on $\mathbb{R}^k$ to a pair of affine connections on a statistical manifold. The α-connections of Section 1.6 interpolate between these two extremes ($\alpha = \pm 1$ recovering the exponential and mixture connections), with the Fisher metric — already identified above as $\nabla^2 A(\boldsymbol{\theta})$ — as the unique Riemannian metric compatible with the entire one-parameter family of connections.

Summary. Maximum entropy subject to linear constraints yields exponential families; the log-partition function A is convex with gradient equal to the mean parameters and Hessian equal to the Fisher information; KL divergence within the family is the Bregman divergence of A; and A's convex conjugate is (minus) the entropy. Four elementary facts, one convex function, and — as Volume III will show — an entire Riemannian geometry.

PART 3

The Asymptotic Equipartition Property

Typical sets, the AEP theorem, source coding via typicality, and connections to statistical mechanics.

3.1 Typical Sequences and the AEP

Asymptotic Equipartition Property (AEP)

Let $X_1, X_2, \ldots$ be i.i.d. with distribution $p(x)$ and entropy $H(X)$. Then

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{\Pr} H(X) \quad \text{as } n \to \infty.$$

Proof. By independence, $p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i)$, so

$$-\frac{1}{n}\log p(X_1, \ldots, X_n) = -\frac{1}{n}\sum_{i=1}^{n} \log p(X_i).$$

The random variables $-\log p(X_i)$ are i.i.d. with mean $\mathbb{E}[-\log p(X)] = H(X)$. By the (weak) law of large numbers, the sample mean converges in probability to $H(X)$.

Typical Set

For $\epsilon > 0$ and block length n, the *ϵ-typical set* is

$$A_\epsilon^{(n)} = \left\{(x_1, \ldots, x_n) \in \mathcal{X}^n : \left|-\frac{1}{n}\log p(x_1, \ldots, x_n) - H(X)\right| < \epsilon\right\}.$$

Equivalently, $x^n \in A_\epsilon^{(n)}$ iff $2^{-n(H+\epsilon)} \le p(x^n) \le 2^{-n(H-\epsilon)}$.

Properties of the typical set.

Property 1: High probability. $\Pr(X^n \in A_\epsilon^{(n)}) \ge 1 - \delta$ for all $\delta > 0$ and sufficiently large n. (Direct consequence of the AEP.)

Property 2: Small size. $|A_\epsilon^{(n)}| \le 2^{n(H+\epsilon)}$. **Proof:** $1 \ge \Pr(A_\epsilon^{(n)}) \ge |A_\epsilon^{(n)}| \cdot 2^{-n(H+\epsilon)}$.

Property 3: Lower bound on size. For large n: $|A_\epsilon^{(n)}| \ge (1-\delta)2^{n(H-\epsilon)}$. **Proof:** $1-\delta \le \Pr(A_\epsilon^{(n)}) \le |A_\epsilon^{(n)}| \cdot 2^{-n(H-\epsilon)}$.

Implication. To leading order, there are 2^{nH} typical sequences, each with probability approximately 2^{-nH}. The probability is essentially equally distributed (equipartitioned) among the typical sequences — hence *asymptotic equipartition property*.

EXAMPLE — Typical sequences of a biased coin

$X \sim \text{Bernoulli}(0.3)$, so $H(X) = h(0.3) = 0.881$ bits. Total sequences of length $n = 20$: $2^{20} = 1{,}048{,}576$. Typical sequences have approximately $0.3 \times 20 = 6$ ones (by the LLN). $|A_\epsilon^{(20)}| \approx 2^{20 \times 0.881} = 2^{17.62} \approx 198{,}000$.

Number of sequences with exactly k ones: $\binom{20}{k}(0.3)^k(0.7)^{20-k}$.

$$k=4: \binom{20}{4}(0.3)^4(0.7)^{16} = 4845 \times 0.0081 \times 0.0023 \approx 0.090 \text{ (fraction of total prob.)}$$

$$k=6: \binom{20}{6}(0.3)^6(0.7)^{14} = 38760 \times 7.3 \times 10^{-4} \times 7.8 \times 10^{-4} \approx 0.192$$

$$k=8: \approx 0.114.$$

The typical set clusters around $k = 5$–7 and contains $\approx 19\%$ of all sequences by number but carries $\approx 92\%$ of the probability.

Strongly typical sets. A stronger notion: x^n is *strongly typical* if the empirical distribution $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[x_i = x]$ satisfies $\|\hat{p} - p\|_1 < \epsilon$. Strong typicality implies weak typicality, and all the properties above hold. Strong typicality is more useful in multi-user information theory because it enables the combinatorial method of counting sequences.

Method of types. The *type* P_{x^n} of a sequence x^n is its empirical distribution. Sequences of the same type form a *type class* $\mathcal{T}(P)$. The number of types for sequences of length n over $|\mathcal{X}|$-ary alphabet is at most $(n+1)^{|\mathcal{X}|}$ (polynomial in n). The size of a type class satisfies

$$\frac{1}{(n+1)^{|\mathcal{X}|}} e^{nH(P)} \leq |\mathcal{T}(P)| \leq e^{nH(P)},$$

and the probability of a type class under distribution q is

$$e^{-n(D_{\mathrm{KL}}(P\,\|\,q)+\delta)} \leq q^n(\mathcal{T}(P)) \leq e^{-nD_{\mathrm{KL}}(P\,\|\,q)}.$$

The method of types is a powerful combinatorial tool for proving coding theorems.

3.2 Source Coding: Definitions and Kraft's Inequality

Lossless Source Code

A *lossless source code* for a random variable X with alphabet $\mathcal{X}$ is a mapping $C : \mathcal{X} \to \{0,1\}^*$ (codewords). The *expected length* is $L(C) = \sum_{x\in\mathcal{X}} p(x)|C(x)|$, where $|C(x)|$ is the length of the codeword for x. A code is *uniquely decodable* if the extension $C^* : \mathcal{X}^* \to \{0,1\}^*$ is injective. A code is *prefix-free* (instantaneous) if no codeword is a prefix of another.

Kraft's Inequality

For any prefix-free binary code with codeword lengths $l_1, l_2, \ldots, l_m$:

$$\sum_{i=1}^{m} 2^{-l_i} \leq 1.$$

Conversely, given lengths satisfying this inequality, a prefix-free code exists.

Proof (necessity). Consider the binary tree of all semi-infinite sequences $\{0,1\}^\infty$. Each codeword of length l_i blocks off a subtree of 2^{-l_i} (as a fraction of the total tree). Prefix-free means these subtrees are disjoint. Since the subtrees are subsets of the unit interval (the total tree), their total measure cannot exceed 1.

Proof (sufficiency). Sort lengths: $l_1 \leq l_2 \leq \ldots \leq l_m$. The i-th codeword is assigned the i-th binary string (in lexicographic order) of the appropriate length that is not a prefix or extension of earlier codewords. The Kraft inequality guarantees this is always possible.

Optimal code lengths. The minimum expected length $L^* = \min_C L(C)$ over prefix-free codes satisfies

$$H(X) \leq L^* < H(X) + 1.$$

Lower bound: by the information inequality and Kraft's inequality (via Lagrange multipliers):

$$L^* = \sum_i p_i l_i \geq \sum_i p_i \log(1/p_i) = H(X).$$

Upper bound: take $l_i = \lceil \log(1/p_i) \rceil$. Then $\sum_i 2^{-l_i} \leq \sum_i p_i = 1$ (Kraft satisfied), and $l_i < \log(1/p_i) + 1$, giving $L^* < H(X) + 1$.

Example — Huffman code and entropy bound

$\mathcal{X} = \{a, b, c, d, e\}$ with probabilities $(0.40, 0.25, 0.15, 0.12, 0.08)$. Huffman code (standard greedy construction):

Symbol	p	Codeword	Length
a	0.40	0	1
b	0.25	10	2
c	0.15	110	3
d	0.12	1110	4
e	0.08	1111	4

$L = 0.40(1) + 0.25(2) + 0.15(3) + 0.12(4) + 0.08(4) = 0.4 + 0.5 + 0.45 + 0.48 + 0.32 = 2.15$ bits. $H(X) = -0.4 \log 0.4 - 0.25 \log 0.25 - 0.15 \log 0.15 - 0.12 \log 0.12 - 0.08 \log 0.08 = 0.529 + 0.5 + 0.411 + 0.362 + 0.292 = 2.094$ bits. Overhead: $2.15 - 2.094 = 0.056$ bits, well within the < 1 bits upper bound.

3.3 Shannon's Source Coding Theorem

Block Source Coding

For source $\{X_i\}$ with entropy rate $\mathcal{H}$, an (n, M, δ) *source code* consists of an encoder $f_n : \mathcal{X}^n \to [M]$ and decoder $g_n : [M] \to \mathcal{X}^n$ achieving $\Pr(g_n(f_n(X^n)) \neq X^n) \leq \delta$. The *rate* is $R = \frac{1}{n} \log M$ bits per source symbol.

Shannon's Source Coding Theorem (Noiseless Channel Coding)

For i.i.d. source X with entropy $H(X)$:

- **Achievability**: For any $R > H(X)$ and $\delta > 0$, there exists n_0 such that for $n \geq n_0$, an $(n, 2^{nR}, \delta)$ lossless source code exists.
- **Converse**: For any $R < H(X)$, any $(n, 2^{nR}, \delta)$ code satisfies $\delta \geq 1 - 2^{-n(H(X)-R)} \to 1$.

Proof of Achievability via Typical Set Coding. Enumerate all sequences in $A_\epsilon^{(n)}$ and assign indices $1, 2, \ldots, |A_\epsilon^{(n)}|$. The encoder maps $x^n \in A_\epsilon^{(n)}$ to its index, and maps $x^n \notin A_\epsilon^{(n)}$ to a fixed sequence (error). The decoder maps index to the corresponding typical sequence. Rate: $\frac{1}{n} \log |A_\epsilon^{(n)}| \leq \frac{1}{n}(n(H + \epsilon)) = H + \epsilon$. By AEP, the probability of error (atypical x^n)

satisfies $\Pr(A_\epsilon^{(n)}) \to 1$, so $\Pr(\text{error}) \to 0$. Setting $\epsilon < R - H(X)$ completes the achievability proof.

Proof of Converse. For a code of rate $R < H(X)$, the decoder's codebook has $M = 2^{nR}$ codewords. The encoder maps $\mathcal{X}^n$ to $[M]$; the decoder maps back. The probability of correct decoding satisfies:

$$\Pr(\text{correct}) = \Pr(g(f(X^n)) = X^n) \leq \Pr(f^{-1}(f(X^n))) \leq \frac{|M|}{|\mathcal{X}^n|}.$$

More precisely, by Fano's inequality applied to the decoder: $H(X^n|f(X^n)) \leq 1 + P_e \cdot nH(X)$. But $H(X^n|f(X^n)) \geq H(X^n) - H(f(X^n)) \geq nH(X) - nR$. Therefore $nH(X) - nR \leq 1 + P_e \cdot nH(X)$, giving $P_e \geq 1 - R/H(X) - 1/(nH(X)) \to 1 - R/H(X) > 0$ as $n \to \infty$.

Example — Source coding for English text

Shannon estimated the entropy rate of English text at approximately $\mathcal{H} \approx 1.0$–1.5 bits per character (via human prediction experiments). Assuming $\mathcal{H} = 1.2$ bits per character: A file of $n = 10^6$ characters requires at least 1.2×10^6 bits $= 150$ KB for lossless compression. A naive 8-bit ASCII encoding uses 10^6 bytes $= 1$ MB. The compression ratio achieved by optimal coding is $8/1.2 \approx 6.67 : 1$. Modern compressors (gzip, bzip2) approach ratios of 3–$5 : 1$ in practice for typical English text, demonstrating the gap between theoretical entropy and achievable practical compression.

3.4 Entropy and Statistical Mechanics: A Bridge

Boltzmann–Gibbs entropy. In statistical mechanics, for a system with microstate space Ω and probability distribution $p(\omega)$ over microstates, the Gibbs entropy is

$$S_G = -k_B \sum_{\omega \in \Omega} p(\omega) \ln p(\omega) = k_B \ln 2 \cdot H(X),$$

where $k_B = 1.38 \times 10^{-23}$ J/K is Boltzmann's constant. At thermal equilibrium, $p(\omega) = e^{-\beta E(\omega)}/Z$ (Boltzmann distribution), where $\beta = 1/(k_B T)$ and $Z = \sum_\omega e^{-\beta E(\omega)}$ is the partition function.

Maximum entropy principle (Jaynes). Jaynes (1957) proposed that the equilibrium distribution in statistical mechanics is simply the distribution that maximises entropy subject to the known macroscopic constraints (energy, particle number, etc.). This identifies the second law not as a separate physical principle but as a consequence of inference: the maximum-entropy distribution represents the most conservative assumption consistent with the available macroscopic information.

Free energy and KL divergence. The Helmholtz free energy $F = -k_B T \ln Z$ satisfies

$$F = \mathbb{E}[E] - TS_G = k_B T D_{\mathrm{KL}}(p \,\|\, \pi)\,,$$

where π is the uniform distribution. Non-equilibrium free energy thus corresponds to KL divergence from the equilibrium distribution.

Example — Maximum entropy for a two-level system

States $\{0, 1\}$ with energies $E_0 = 0$, $E_1 = \epsilon$. Constraint: $\mathbb{E}[E] = \bar{E}$. Maximise $H(p_0, p_1) = h(p_1)$ subject to $p_1 \epsilon = \bar{E}$ and $p_0 + p_1 = 1$. Solution: $p_1 = \bar{E}/\epsilon$ (forced by constraint), $p_0 = 1 - \bar{E}/\epsilon$. $H = h(\bar{E}/\epsilon)$. The Boltzmann distribution gives

$p_1 = e^{-\beta\epsilon}/(1+e^{-\beta\epsilon})$, so β is determined by $\bar{E} = \epsilon e^{-\beta\epsilon}/(1+e^{-\beta\epsilon})$. As $T \to \infty$ ($\beta \to 0$): $p_1 \to 1/2$, $H \to 1$ bits (maximum entropy). As $T \to 0$ ($\beta \to \infty$): $p_1 \to 0$, $H \to 0$ (ground state, no uncertainty).

3.5 Universal Source Coding and the LZ Algorithm

The problem of universality. Shannon's source coding theorem assumes the source distribution p is known. In practice, p is unknown; a *universal* source code must compress *any* source at the optimal rate without knowing p.

Lempel–Ziv (LZ) compression. The LZ78 (Ziv–Lempel 1978) algorithm parses the sequence $x_1 x_2 \ldots x_n$ greedily into distinct phrases: $x_1, x_2, \ldots$ until a new symbol creates a phrase not seen before; record the index of the longest previously seen phrase and the new symbol. LZ77 uses a sliding window variant.

Universality theorem (Ziv–Lempel). For any stationary ergodic source with entropy rate $\mathcal{H}$:

$$\limsup_{n\to\infty} \frac{L_{\mathrm{LZ}}(X_1^n)}{n} \leq \mathcal{H} \quad \text{a.s.}$$

That is, LZ78 achieves the entropy rate almost surely, without knowing the source. This is remarkable: the algorithm is simultaneously optimal for all stationary ergodic sources.

Convergence rate. The redundancy (excess rate above $\mathcal{H}$) for LZ78 converges as $O(\log\log n/\log n)$, slower than the $O(1/n)$ for codes tuned to the known source but universal.

Example — LZ78 parsing of a simple sequence

$x^{10} = 0100101011$. Parsing greedily into distinct phrases:

1. 0 (new)
2. 1 (new)
3. 00 ("0" seen; "00" new)
4. 10 ("1" seen; "10" new)
5. 101 ("10" seen; "101" new)
6. $1 \to$ remainder.

5 complete phrases from 10 symbols. Code each phrase as (index of prefix, new symbol): phrase 1: $(0,0)$, phrase 2: $(0,1)$, phrase 3: $(1,0)$, phrase 4: $(2,0)$, phrase 5: $(4,1)$. Compressed representation: $\lceil \log 4 \rceil = 2$ bits for index plus 1 bit for symbol $= 3$ bits per phrase, total $5 \times 3 = 15$ bits for 10 symbols, rate $= 1.5$ bits/symbol. The source entropy $H = h(0.5) = 1$ bit/symbol; the redundancy (0.5 bits/symbol) decreases as n grows.

3.6 The Method of Types in Detail

The AEP of Section 3.1 classified sequences into two camps — typical and atypical — and showed that the typical set carries essentially all the probability mass while containing roughly $2^{nH(X)}$ sequences, each of probability roughly $2^{-nH(X)}$. The *method of types* refines this picture dramatically: instead of a single typical set, it partitions the entire sample space $\mathcal{X}^n$ into *type classes* indexed by empirical distributions, and gives an exact (not merely asymptotic) formula for the probability and size of every type class. This refinement is the single most versatile combinatorial tool in information theory: it underlies the proof of Sanov's theorem

on large deviations (previewed in Appendix D), gives sharp non-asymptotic bounds for hypothesis testing, and provides the cleanest possible construction of universal (distribution-free) source codes.

Type of a Sequence

For a sequence $x^n = (x_1, \dots, x_n) \in \mathcal{X}^n$, the *type* (or empirical distribution) P_{x^n} is the probability distribution on $\mathcal{X}$ defined by

$$P_{x^n}(a) = \frac{1}{n}\big|\{i : x_i = a\}\big|, \qquad a \in \mathcal{X},$$

i.e. the relative frequency of each symbol in x^n. The set of all possible types of sequences of length n over $\mathcal{X}$ is denoted $\mathcal{P}_n$, and the *type class* of a type $Q \in \mathcal{P}_n$ is

$$T(Q) = \{x^n \in \mathcal{X}^n : P_{x^n} = Q\}.$$

Counting types and type classes. A type $Q \in \mathcal{P}_n$ is specified by the counts $n_a = nQ(a)$ for each $a \in \mathcal{X}$, which are non-negative integers summing to n. The number of such integer vectors is $\binom{n+|\mathcal{X}|-1}{|\mathcal{X}|-1}$ (stars-and-bars), which grows only *polynomially* in n — specifically as $O(n^{|\mathcal{X}|-1})$ — while $|\mathcal{X}|^n$ grows exponentially. Therefore:

$$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|},$$

a crude but extremely useful bound: *there are only polynomially many types*, so by the pigeonhole principle, at least one type class must contain an exponentially large fraction of all sequences — and indeed, as we will see, the type class of the "true" distribution p contains *almost all* the probability mass under p^n.

Exact Probability of a Type Class

For $x^n \in T(Q)$ and p the true (i.i.d.) generating distribution,

$$p^n(x^n) = \prod_{i=1}^{n} p(x_i) = \prod_{a \in \mathcal{X}} p(a)^{nQ(a)} = 2^{-n(H(Q)+D_{\mathrm{KL}}(Q\,\|\,p))}.$$

Derivation. Taking logarithms (base 2),

$$\log p^n(x^n) = \sum_a nQ(a)\log p(a) = n\sum_a Q(a)\log p(a).$$

Now write $\log p(a) = \log Q(a) - \log\frac{Q(a)}{p(a)}$ (valid wherever $Q(a) > 0$, with the convention $0\log\frac{0}{p(a)} = 0$). Then

$$\sum_a Q(a)\log p(a) = \sum_a Q(a)\log Q(a) - \sum_a Q(a)\log\frac{Q(a)}{p(a)} = -H(Q) - D_{\mathrm{KL}}(Q\,\|\,p)\,.$$

Substituting back, $\log p^n(x^n) = -n(H(Q) + D_{\mathrm{KL}}(Q\,\|\,p))$, i.e. $p^n(x^n) = 2^{-n(H(Q)+D_{\mathrm{KL}}(Q\,\|\,p))}$.

The crucial observation. The probability $p^n(x^n)$ depends on x^n *only through its type* $Q = P_{x^n}$ — every sequence of the same type is equally likely under p^n, regardless of the order of symbols. This is, of course, simply the statement that p^n is an exchangeable (i.i.d.) measure, but the method of types makes the consequence explicit and quantitative: *the entire probabilistic content of a length-n i.i.d. sample reduces to two numbers, $H(Q)$ and $D_{\mathrm{KL}}(Q\,\|\,p)$*, both computable from the empirical distribution Q alone.

Size of a Type Class

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(Q)} \le |T(Q)| \le 2^{nH(Q)}.$$

Derivation of the upper bound. Since $T(Q)$ is a subset of $\mathcal{X}^n$ and every sequence in $T(Q)$ has the same probability $Q^n(x^n) = 2^{-nH(Q)}$ under the measure Q^n (apply the previous lemma with $p = Q$, so $D_{\mathrm{KL}}(Q \,\|\, Q) = 0$), and probabilities sum to at most 1,

$$1 \ge \sum_{x^n \in T(Q)} Q^n(x^n) = |T(Q)| \cdot 2^{-nH(Q)} \implies |T(Q)| \le 2^{nH(Q)}.$$

Derivation of the lower bound. The lower bound requires showing that $T(Q)$ is, among all type classes, not too much smaller than its trivial upper bound — intuitively, $T(Q)$ is the *largest* type class for the distribution Q itself (it is, after all, the "typical set" for i.i.d. Q^n-samples), and the total probability $\sum_{x^n \in T(Q)} Q^n(x^n)$ cannot be too small compared to $1/|\mathcal{P}_n|$ because $T(Q)$ is the most probable type class under Q^n. A short argument (comparing $Q^n(T(Q))$ to $Q^n(T(Q'))$ for any other type Q' via a ratio bound shows $Q^n(T(Q)) \ge Q^n(T(Q'))$ for all $Q' \in \mathcal{P}_n$, hence $Q^n(T(Q)) \ge 1/|\mathcal{P}_n| \ge (n+1)^{-|\mathcal{X}|}$ since the type classes partition $\mathcal{X}^n$ and there are at most $(n+1)^{|\mathcal{X}|}$ of them) combined with $Q^n(T(Q)) = |T(Q)|2^{-nH(Q)}$ from above gives $|T(Q)|2^{-nH(Q)} \ge (n+1)^{-|\mathcal{X}|}$, i.e. $|T(Q)| \ge (n+1)^{-|\mathcal{X}|} 2^{nH(Q)}$.

Combining the two lemmas: probability of a type class.

$$p^n(T(Q)) = |T(Q)| \cdot p^n(x^n)\Big|_{x^n \in T(Q)} \in \Big[(n+1)^{-|\mathcal{X}|},\, 1\Big] \cdot 2^{nH(Q)} \cdot 2^{-n(H(Q)+D_{\mathrm{KL}}(Q\,\|\,p))} = \Big[(n+1)^{-|\mathcal{X}|},\, 1\Big] \cdot 2^{-nD_{\mathrm{KL}}(Q\,\|\,p)}.$$

Hence, up to the polynomial factor $(n+1)^{|\mathcal{X}|}$ (which is sub-exponential and therefore negligible on the scale of the exponent),

$$p^n(T(Q)) \approx 2^{-nD_{\mathrm{KL}}(Q\,\|\,p)}.$$

This single approximate formula is the heart of the method of types: *the probability that an i.i.d. p-sample has empirical distribution Q decays exponentially in n at rate $D_{\mathrm{KL}}(Q \,\|\, p)$* — the KL divergence between the observed empirical distribution and the true one. The closer Q is to p (in KL divergence), the more likely it is to be observed; $Q = p$ gives $D_{\mathrm{KL}}(p \,\|\, p) = 0$, i.e. probability $\approx 2^{-0} = 1$ (up to the polynomial factor) — consistent with the law of large numbers, which says the empirical distribution converges to p.

Example — Method of types for a biased coin, $n = 4$

Let $\mathcal{X} = \{0, 1\}$, $p = (0.7, 0.3)$ (so $p(1) = 0.7$ is the probability of a "success"), and $n = 4$.
Possible types: $Q_k = (1 - k/4,\, k/4)$ for $k = 0, 1, 2, 3, 4$, i.e. $Q_0 = (1, 0)$, $Q_1 = (0.75, 0.25)$, $Q_2 = (0.5, 0.5)$, $Q_3 = (0.25, 0.75)$, $Q_4 = (0, 1)$. There are $|\mathcal{P}_4| = 5$ types, matching $\binom{4+1}{1} = 5$ from stars-and-bars (with $|\mathcal{X}| - 1 = 1$).
Type class sizes: $|T(Q_k)| = \binom{4}{k}$, i.e. $1, 4, 6, 4, 1$ for $k = 0, \dots, 4$, summing to $2^4 = 16$ as required.
$H(Q_k)$ **values (bits):** $H(Q_0) = H(Q_4) = 0$; $H(Q_1) = H(Q_3) = h(0.25) = 0.8113$; $H(Q_2) = h(0.5) = 1$.
Check of the size bounds for Q_2: $2^{nH(Q_2)} = 2^{4\times 1} = 16$. The lower bound $(n+1)^{-|\mathcal{X}|}2^{nH(Q)} = (5)^{-2} \times 16 = 16/25 = 0.64$. Indeed $|T(Q_2)| = 6$ lies in $[0.64, 16]$ — comfortably, since $n = 4$ is small and the bounds are loose for small n (they become

tight only in relative-exponent terms as $n \to \infty$).
$D_{\mathrm{KL}}(Q_k \,\|\, p)$ **values (bits),** $p = (0.7, 0.3)$**:**

$$D_{\mathrm{KL}}(Q_0 \,\|\, p) = 1 \log_2 \frac{1}{0.7} + 0 = 0.5146,$$
$$D_{\mathrm{KL}}(Q_1 \,\|\, p) = 0.75 \log_2 \frac{0.75}{0.7} + 0.25 \log_2 \frac{0.25}{0.3} = 0.75(0.1024) + 0.25(-0.2630) = 0.0768 - 0.0657 = 0.0111,$$
$$D_{\mathrm{KL}}(Q_2 \,\|\, p) = 0.5 \log_2 \frac{0.5}{0.7} + 0.5 \log_2 \frac{0.5}{0.3} = 0.5(-0.4854) + 0.5(0.7370) = -0.2427 + 0.3685 = 0.1258,$$
$$D_{\mathrm{KL}}(Q_3 \,\|\, p) = 0.25 \log_2 \frac{0.25}{0.7} + 0.75 \log_2 \frac{0.75}{0.3} = 0.25(-1.4854) + 0.75(1.3219) = -0.3714 + 0.9914 = 0.620$$
$$D_{\mathrm{KL}}(Q_4 \,\|\, p) = 1 \log_2 \frac{1}{0.3} + 0 = 1.7370.$$

Exact probabilities $p^4(T(Q_k)) = \binom{4}{k} 0.7^{4-k} 0.3^k$**:**

$$k = 0 : 1 \times 0.2401 = 0.2401, \quad k = 1 : \ 4 \times 0.7^3 \times 0.3 = 4 \times 0.343 \times 0.3 = 0.4116,$$
$$k = 2 : 6 \times 0.49 \times 0.09 = 0.2646, \quad k = 3 : \ 4 \times 0.7 \times 0.027 = 0.0756, \quad k = 4 : \ 0.0081.$$

(Sum: $0.2401 + 0.4116 + 0.2646 + 0.0756 + 0.0081 = 1.0000$, as required.)
Cross-check via the exponential approximation $2^{-n D_{\mathrm{KL}}(Q_k \,\|\, p)}$**:** $k = 1$ (closest type to $p = 0.3$): $2^{-4 \times 0.0111} = 2^{-0.0444} = 0.970$, compared to the exact 0.4116 — the approximation $2^{-n D_{\mathrm{KL}}(\,\|\,)}$ is off by the polynomial/combinatorial prefactor $|T(Q_1)| 2^{-n H(Q_1)} = 4 \times 2^{-4 \times 0.8113} = 4 \times 2^{-3.245} = 4 \times 0.1056 = 0.4225$, and indeed $0.4225 \times 0.970 \approx 0.4098 \approx 0.4116$ (small rounding). The exact value is the product of the (sub-exponential) prefactor and the (exponential) $2^{-n D_{\mathrm{KL}}(Q \,\|\, p)}$ term — at $n = 4$ the prefactor is not negligible, but as $n \to \infty$ the exponential term dominates and $-\frac{1}{n} \log p^n(T(Q)) \to D_{\mathrm{KL}}(Q \,\|\, p)$ exactly.

Sanov's theorem. Let $E \subseteq \mathcal{P}$ be a set of distributions on $\mathcal{X}$ (thought of as a "large deviation event" — e.g. $E = \{Q : Q(1) \geq 0.5\}$ for the biased-coin example, an event that is *atypical* if $p(1) = 0.3 < 0.5$). Then

$$p^n(\{x^n : P_{x^n} \in E\}) \leq (n+1)^{|\mathcal{X}|} \, 2^{-n \, D_{\mathrm{KL}}(Q^* \,\|\, p)}, \qquad Q^* = \operatorname*{arg\,min}_{Q \in E \cap \mathcal{P}_n} D_{\mathrm{KL}}(Q \,\|\, p) \,,$$

and, if E is the closure of its interior (a mild regularity condition satisfied by all the sets of practical interest, such as half-spaces defined by linear inequalities on Q),

$$\lim_{n \to \infty} -\frac{1}{n} \log p^n(\{x^n : P_{x^n} \in E\}) = \min_{Q \in E} D_{\mathrm{KL}}(Q \,\|\, p) \,.$$

Proof of the upper bound. $\{x^n : P_{x^n} \in E\} = \bigcup_{Q \in E \cap \mathcal{P}_n} T(Q)$, a union over (at most $(n+1)^{|\mathcal{X}|}$) type classes. By the union bound and $p^n(T(Q)) \leq 2^{-n D_{\mathrm{KL}}(Q \,\|\, p)}$ (dropping the polynomial lower-bound factor, which only helps the upper bound),

$$p^n(\{x^n : P_{x^n} \in E\}) \leq \sum_{Q \in E \cap \mathcal{P}_n} 2^{-n D_{\mathrm{KL}}(Q \,\|\, p)} \leq (n+1)^{|\mathcal{X}|} \max_{Q \in E \cap \mathcal{P}_n} 2^{-n D_{\mathrm{KL}}(Q \,\|\, p)} = (n+1)^{|\mathcal{X}|} 2^{-n D_{\mathrm{KL}}(Q^* \,\|\, p)},$$

where the middle inequality bounds the sum of at most $(n+1)^{|\mathcal{X}|}$ terms by $(n+1)^{|\mathcal{X}|}$ times the largest term (smallest exponent $D_{\mathrm{KL}}(Q \,\|\, p)$, hence largest $2^{-n D_{\mathrm{KL}}(Q \,\|\, p)}$).

The lower bound and the exponential rate. The matching lower bound (up to the same polynomial factor, which vanishes in the $-\frac{1}{n} \log(\cdot)$ limit) follows by restricting attention to

the single type class $T(Q^*)$ nearest to E and applying the type-class lower bound $p^n(T(Q^*)) \geq (n+1)^{-|\mathcal{X}|}2^{-nD_{\mathrm{KL}}(Q^*\,\|\,p)}$. Combining upper and lower bounds and taking $-\frac{1}{n}\log(\cdot)$, the polynomial prefactors $(n+1)^{\pm|\mathcal{X}|}$ contribute $O(\log n/n) \to 0$, leaving exactly $D_{\mathrm{KL}}(Q^*\,\|\,p)$.

Interpretation: Sanov's theorem as the entropy-side of large deviations. Sanov's theorem is the multivariate, distribution-valued generalization of Cramér's theorem on large deviations for sums of i.i.d. random variables. Cramér's theorem says $\Pr(\bar{X}_n \geq a) \approx 2^{-nI(a)}$ for a above the mean, where I is a rate function given by a Legendre transform of the cumulant generating function; Sanov's theorem says $\Pr(P_{X^n} \in E) \approx 2^{-nD_{\mathrm{KL}}(Q^*\,\|\,p)}$ for E a set of "atypical" empirical distributions, with rate function $D_{\mathrm{KL}}(\cdot\,\|\,p)$. In fact Cramér's theorem can be *derived* from Sanov's theorem by taking $E = \{Q : \mathbb{E}_Q[X] \geq a\}$ and observing that the KL-projection Q^* of p onto this half-space is exactly the exponentially-tilted distribution $Q^*(x) \propto p(x)e^{\lambda x}$ that appears in the Chernoff-bound / Legendre-transform proof of Cramér's theorem — the tilting parameter λ is precisely the Lagrange multiplier for the constraint $\mathbb{E}_Q[X] = a$ in the KL-minimization $\min_{Q:\mathbb{E}_Q[X]=a} D_{\mathrm{KL}}(Q\,\|\,p)$. This connection between information projections (KL-minimization over constraint sets) and exponential tilting is a recurring theme that will be developed fully in Volume III under the heading of *dual foliations* of the statistical manifold.

Universal source coding via types: the two-stage code. The method of types also yields, with no further machinery, a simple *universal* source code — one that does not need to know p in advance, yet achieves a redundancy (excess code length over the entropy bound) that vanishes as $n \to \infty$.

Construction. To encode $x^n \in \mathcal{X}^n$: first encode its type $Q = P_{x^n}$ using $\lceil |\mathcal{X}| \log(n+1) \rceil$ bits (there are at most $(n+1)^{|\mathcal{X}|}$ types, so this many bits suffice to index any of them); then, given the decoder now knows Q, encode *which* sequence within $T(Q)$ is x^n, using $\lceil \log |T(Q)| \rceil \leq nH(Q)$ bits (a fixed-length code over the $|T(Q)|$ equiprobable-given-Q sequences in the type class). Total length:

$$\ell(x^n) \leq nH(P_{x^n}) + |\mathcal{X}| \log(n+1) + O(1).$$

Redundancy analysis. The expected code length under the true p is $\mathbb{E}_p[\ell(X^n)] \leq n\mathbb{E}_p[H(P_{X^n})] + |\mathcal{X}| \log(n+1) + O(1)$. Since $H(P_{x^n}) \leq H(p)$ is *not* generally true pointwise (the empirical entropy can exceed or be below the true entropy for a particular sample), but $\mathbb{E}_p[H(P_{X^n})] \to H(p)$ as $n \to \infty$ by the law of large numbers (continuity of H) together with a uniform integrability argument, we get

$$\frac{1}{n}\mathbb{E}_p[\ell(X^n)] \leq H(p) + \frac{|\mathcal{X}| \log(n+1)}{n} + o(1) \xrightarrow{n\to\infty} H(p).$$

The *redundancy* — the excess rate over the entropy $H(p)$ — is $O\big(\frac{|\mathcal{X}| \log n}{n}\big)$, decaying to zero. Crucially, this holds for *every* p simultaneously: the code construction above never referenced p at all, only n and $|\mathcal{X}|$. This is the simplest example of a *universal code* — a single code that is asymptotically optimal for every source in a class (here, all i.i.d. sources on $\mathcal{X}$), foreshadowing the Lempel–Ziv-type universal codes of Section 3.5, which achieve the same vanishing-redundancy property for the much larger class of all stationary ergodic sources, at the cost of a more intricate (dictionary-based) construction.

Minimax redundancy and the type-counting bound. The two-stage code above has redundancy $O(|\mathcal{X}| \log n/n)$ per symbol, i.e. total redundancy $O(|\mathcal{X}| \log n)$ over the whole block of length n. A deeper result (Shtarkov, Krichevsky–Trofimov) shows that this is essentially *optimal*: the minimax redundancy over all codes and all i.i.d. sources on an alphabet of size $|\mathcal{X}|$ is

$$R_n^* = \frac{|\mathcal{X}| - 1}{2} \log n + O(1),$$

matching the two-stage bound up to the constant in front of $\log n$ (the factor $\frac{|\mathcal{X}|-1}{2}$ versus $|\mathcal{X}|$ reflects that the two-stage code's type-encoding stage is itself slightly wasteful — a more careful code that assigns shorter codewords to more probable types, rather than a uniform-length type index, recovers the sharper constant). The quantity $\frac{|\mathcal{X}|-1}{2}\log n$ has a clean interpretation: it is (twice) the log of the "volume," in the Fisher-information metric, of the $(|\mathcal{X}|-1)$-dimensional simplex of possible types at resolution $1/n$ — a first hint of the deep connection between universal coding, Bayesian model selection (the Bayesian Information Criterion's $\frac{k}{2}\log n$ penalty has exactly this form, with $k = |\mathcal{X}|-1$ the number of free parameters), and the Riemannian volume form induced by the Fisher metric, the central object of Volume III.

Looking ahead. The method of types will reappear in Volume II as the combinatorial foundation for the error-exponent analysis of channel codes (where "types" of *joint* sequences (x^n, y^n) — the joint empirical distribution of input and output — play the role that types of single sequences play here), and the KL-divergence rate function $D_{\mathrm{KL}}(Q\,\|\,p)$ derived in Sanov's theorem is exactly the exponent governing the probability of decoding error in the random-coding bound, linking the source-coding combinatorics of this section directly to the channel-coding theorems of Part 4.

3.7 Hypothesis Testing, Stein's Lemma, and the Chernoff Information

The method of types (Section 3.6) showed that the probability of observing an empirical distribution Q under an i.i.d. source p decays as $2^{-nD_{\mathrm{KL}}(Q\,\|\,p)}$. This section applies that single fact to the oldest problem in mathematical statistics — deciding, from n observations, which of two candidate distributions generated the data — and shows that the fundamental error exponents of hypothesis testing are *literally* KL divergences and Rényi divergences. The results obtained here (Stein's lemma and the Chernoff–Stein lemma) are not mere analogies to the channel-coding error exponents of Section 4.5: they are the same mathematical objects, and the connection becomes a precise duality once the channel coding problem is itself recast as a hypothesis-testing problem (the meta-converse framework mentioned at the end of Part 4 makes this duality the organizing principle of finite-blocklength information theory).

Binary Hypothesis Testing

Given i.i.d. observations $X^n = (X_1, \dots, X_n)$, consider two hypotheses about their common distribution: $H_0 : X_i \sim p$ versus $H_1 : X_i \sim q$. A *test* is a partition of $\mathcal{X}^n$ into an acceptance region A_n (decide H_0) and its complement A_n^c (decide H_1). The two error probabilities are

$$\alpha_n = \Pr_p(X^n \in A_n^c) \quad \text{(type-I error: reject } H_0 \text{ when true)},$$

$$\beta_n = \Pr_q(X^n \in A_n) \quad \text{(type-II error: accept } H_0 \text{ when } H_1 \text{ true)}.$$

The Neyman–Pearson lemma. Among all tests with $\alpha_n \le \epsilon$, the test that minimizes β_n is a *likelihood-ratio test*: $A_n = \{x^n : p(x^n)/q(x^n) > \tau\}$ for an appropriately chosen threshold τ (with possible randomization on the boundary $p(x^n)/q(x^n) = \tau$ to hit $\alpha_n = \epsilon$ exactly). *Proof sketch.* For any other test with region B_n satisfying $\Pr_p(B_n^c) \le \epsilon = \Pr_p(A_n^c)$, write $\beta_n(A_n) - \beta_n(B_n) = \Pr_q(A_n) - \Pr_q(B_n) = \Pr_q(A_n \setminus B_n) - \Pr_q(B_n \setminus A_n)$. On $A_n \setminus B_n \subseteq A_n$, $p(x^n)/q(x^n) > \tau$ so $q(x^n) < p(x^n)/\tau$; on $B_n \setminus A_n \subseteq A_n^c$, $p(x^n)/q(x^n) \le \tau$ so $q(x^n) \ge p(x^n)/\tau$. Hence $\Pr_q(A_n \setminus B_n) < \frac{1}{\tau}\Pr_p(A_n \setminus B_n)$ and $\Pr_q(B_n \setminus A_n) \ge \frac{1}{\tau}\Pr_p(B_n \setminus A_n)$, and since $\Pr_p(A_n \setminus B_n) - \Pr_p(B_n \setminus A_n) = \Pr_p(A_n) - \Pr_p(B_n) = \Pr_p(B_n^c) - \Pr_p(A_n^c) \ge 0$ (by the

constraint), combining the inequalities gives $\beta_n(A_n) \le \beta_n(B_n)$.

From exact optimality to exponential rates. The Neyman–Pearson lemma characterizes the *exactly* optimal test for every finite n, but the resulting β_n has no simple closed form in general. The method of types instead asks for the *exponential rate* at which the optimal β_n decays as $n \to \infty$, for a *fixed* (non-vanishing) bound ϵ on α_n — and this rate *does* have a simple closed form.

Stein's Lemma

Fix $\epsilon \in (0,1)$. Let $\beta_n^*(\epsilon) = \min\{\beta_n : \alpha_n \le \epsilon\}$ be the smallest type-II error achievable by any test with type-I error at most ϵ. Then

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{1}{\beta_n^*(\epsilon)} = D_{\mathrm{KL}}(p \,\|\, q)\,,$$

independent of $\epsilon \in (0,1)$.

Proof via the method of types: achievability. Consider the test $A_n = \{x^n : D_{\mathrm{KL}}(P_{x^n} \,\|\, p) \le \delta\}$ for small $\delta > 0$ — accept H_0 iff the empirical distribution of x^n is within KL-divergence δ of p. By the AEP (Section 3.1) / the law of large numbers for the empirical distribution, $\Pr_p(P_{X^n} \to p) = 1$, so $\Pr_p(D_{\mathrm{KL}}(P_{X^n} \,\|\, p) > \delta) \to 0$, giving $\alpha_n \to 0 \le \epsilon$ for any $\epsilon > 0$ and n large enough. Now bound $\beta_n = \Pr_q(A_n) = \Pr_q(D_{\mathrm{KL}}(P_{X^n} \,\|\, p) \le \delta) = \sum_{Q : D_{\mathrm{KL}}(Q\,\|\,p) \le \delta} q^n(T(Q))$. By the type-class probability bound, $q^n(T(Q)) \le 2^{-nD_{\mathrm{KL}}(Q\,\|\,q)}$, so

$$\beta_n \le (n+1)^{|\mathcal{X}|} \max_{Q : D_{\mathrm{KL}}(Q\,\|\,p) \le \delta} 2^{-nD_{\mathrm{KL}}(Q\,\|\,q)}.$$

As $\delta \to 0$, the maximizing $Q \to p$ (continuity of $D_{\mathrm{KL}}(\cdot \,\|\, q)$), so $\max_{Q : D_{\mathrm{KL}}(Q\,\|\,p) \le \delta} D_{\mathrm{KL}}(Q \,\|\, q) \to D_{\mathrm{KL}}(p \,\|\, q)$, giving $-\frac{1}{n}\log\beta_n \ge D_{\mathrm{KL}}(p \,\|\, q) - \eta(\delta)$ for $\eta(\delta) \to 0$, i.e. $\liminf_n -\frac{1}{n}\log\beta_n^*(\epsilon) \ge D_{\mathrm{KL}}(p \,\|\, q)$ after taking $\delta \to 0$.

Proof via the method of types: converse. For *any* test with $\alpha_n \le \epsilon < 1$, we must show $-\frac{1}{n}\log\beta_n \le D_{\mathrm{KL}}(p \,\|\, q) + o(1)$. Suppose for contradiction that some test achieves $\beta_n \le 2^{-n(D_{\mathrm{KL}}(p\,\|\,q)+2\eta)}$ for some $\eta > 0$ and infinitely many n. Decompose $A_n = \bigcup_Q A_n \cap T(Q)$ over types Q. For types Q with $D_{\mathrm{KL}}(Q \,\|\, p) \le \eta$ (a "typical" set of types, by the same continuity argument), $p^n(T(Q))$ is close to 1 in aggregate (the union of such type classes has p^n-probability $\to 1$ by the law of large numbers), so $\alpha_n \le \epsilon$ forces A_n to contain "most" of $\bigcup_{Q : D_{\mathrm{KL}}(Q\,\|\,p) \le \eta} T(Q)$ (else $\alpha_n = \Pr_p(A_n^c)$ would be too large). But for such Q, $D_{\mathrm{KL}}(Q \,\|\, q) \ge D_{\mathrm{KL}}(p \,\|\, q) - O(\eta)$ (reverse triangle-inequality-type continuity), so $q^n(T(Q)) \ge (n+1)^{-|\mathcal{X}|} 2^{-nD_{\mathrm{KL}}(Q\,\|\,q)} \ge (n+1)^{-|\mathcal{X}|} 2^{-n(D_{\mathrm{KL}}(p\,\|\,q)+O(\eta))}$, and since A_n contains most of $T(Q)$, $\beta_n = q^n(A_n) \gtrsim 2^{-n(D_{\mathrm{KL}}(p\,\|\,q)+O(\eta))}$, contradicting $\beta_n \le 2^{-n(D_{\mathrm{KL}}(p\,\|\,q)+2\eta)}$ for η small enough. Hence $\limsup_n -\frac{1}{n}\log\beta_n^*(\epsilon) \le D_{\mathrm{KL}}(p \,\|\, q)$. Combined with the achievability direction, the limit exists and equals $D_{\mathrm{KL}}(p \,\|\, q)$ exactly, for every $\epsilon \in (0,1)$.

Interpretation: the asymmetry of Stein's lemma. Stein's lemma is strikingly asymmetric: the exponent is $D_{\mathrm{KL}}(p \,\|\, q)$ — the divergence *from the true distribution* p *(under* H_0*) to the alternative* q — and does not depend on which error probability (α_n or β_n) is held fixed at ϵ versus driven to zero, as long as one of them is fixed at a constant $\in (0,1)$ and the other is exponentially minimized. This asymmetry mirrors the asymmetry of KL divergence itself (Section 2.5): $D_{\mathrm{KL}}(p \,\|\, q) \ne D_{\mathrm{KL}}(q \,\|\, p)$ in general, and Stein's lemma tells us that H_0 (the hypothesis whose distribution appears as the *first* argument of the KL divergence achieved) is the one being protected by the constraint $\alpha_n \le \epsilon$, while H_1's error probability β_n is the one being exponentially crushed.

Example — Stein exponent for two Bernoulli hypotheses

$H_0 : p = \text{Bernoulli}(0.5)$, $H_1 : q = \text{Bernoulli}(0.7)$, base-2 logarithms.
Stein exponent (testing H_0 vs. H_1, fixed $\alpha_n = \epsilon$):

$$D_{\mathrm{KL}}(p \,\|\, q) = 0.5 \log_2 \frac{0.5}{0.7} + 0.5 \log_2 \frac{0.5}{0.3} = 0.5(-0.485) + 0.5(0.737) = 0.126 \text{ bits}.$$

So $\beta_n^*(\epsilon) \approx 2^{-0.126n}$: to drive the type-II error below, say, 2^{-10} requires roughly $n \approx 10/0.126 \approx 79$ samples, for any fixed $\alpha_n = \epsilon \in (0,1)$ — even $\epsilon = 0.4999$, barely better than a coin flip on the type-I side, gives the *same* asymptotic exponent. This is the sense in which Stein's lemma says the type-I error constraint is "free" asymptotically: any fixed, non-vanishing tolerance for α_n yields the same exponential rate for β_n.
Reversed roles (testing with $D_{\mathrm{KL}}(q \,\|\, p)$): if instead the labeling of hypotheses is reversed — $H_0 : q$, $H_1 : p$ — so that β_n is exponentially minimized while α_n is held at a fixed tolerance, the relevant exponent becomes $D_{\mathrm{KL}}(q \,\|\, p)$:

$$D_{\mathrm{KL}}(q \,\|\, p) = \sum_x q(x) \log_2 \frac{q(x)}{p(x)} = 0.7 \log_2 \frac{0.7}{0.5} + 0.3 \log_2 \frac{0.3}{0.5} = 0.7(0.485) + 0.3(-0.737) = 0.3395 - 0.2211 =$$

Both exponents are positive, as required, but unequal — $D_{\mathrm{KL}}(p \,\|\, q) = 0.126$ bits versus $D_{\mathrm{KL}}(q \,\|\, p) = 0.1184$ bits — a concrete illustration that *which* hypothesis is designated H_0 (and thus protected by the $\alpha_n \le \epsilon$ constraint) determines which of the two, generally distinct, KL divergences governs the achievable exponential rate.

Symmetric hypothesis testing: the Chernoff–Stein lemma. Stein's lemma treats H_0 and H_1 asymmetrically, which is appropriate when one hypothesis (say, "no effect") deserves protection (as in classical significance testing). When both error types are equally important — e.g. in a symmetric Bayesian decision problem with equal priors — the relevant quantity is the exponent of the *total* (or *Bayes*) error probability $P_e^{(n)} = \pi_0 \alpha_n + \pi_1 \beta_n$ for a test chosen to minimize $P_e^{(n)}$.

Chernoff Information

$$C(p, q) = -\min_{0 \le \lambda \le 1} \log \left(\sum_x p(x)^\lambda q(x)^{1-\lambda} \right).$$

The Chernoff–Stein lemma. For any priors $\pi_0, \pi_1 \in (0,1)$, the Bayes-optimal test (a likelihood-ratio test with threshold $\tau = \pi_1/\pi_0$, by Neyman–Pearson applied to the Bayes risk) satisfies

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{1}{P_e^{*(n)}} = C(p, q),$$

independent of $\pi_0, \pi_1 \in (0,1)$ (the priors affect only the sub-exponential prefactor, not the exponential rate).

Connection to Rényi divergence. Recall from Section 1.6 that the Rényi divergence of order α is $D_\alpha(p\|q) = \frac{1}{\alpha - 1} \log \sum_x p(x)^\alpha q(x)^{1-\alpha}$. The quantity inside the Chernoff information, $\log \sum_x p(x)^\lambda q(x)^{1-\lambda}$, is therefore $(\lambda - 1) D_\lambda(p\|q)$. For $\lambda \in (0,1)$, $\lambda - 1 < 0$, so $\log \sum_x p(x)^\lambda q(x)^{1-\lambda} = (\lambda - 1) D_\lambda(p\|q) \le 0$ (since $D_\lambda \ge 0$), confirming that the quantity being minimized (then negated) in the definition of $C(p, q)$ is indeed non-negative, as required for an error exponent. *The Chernoff information is the Rényi divergence $D_\lambda(p\|q)$, scaled by $(1 - \lambda)$ and optimized*

over $\lambda \in (0,1)$:

$$C(p,q) = \max_{0\le\lambda\le 1} (1-\lambda) D_\lambda(p\|q).$$

This identity is the precise sense in which the Rényi divergence family of Section 1.6 is not merely a mathematical curiosity but the *exact* answer to an operational question: each value of λ corresponds to a particular "tilted" large-deviations exponent, and the Chernoff information picks out the value of λ — equivalently, the particular tilted distribution $Q_\lambda(x) \propto p(x)^\lambda q(x)^{1-\lambda}$, the same exponential-tilting construction encountered in the Sanov's-theorem/Cramér's-theorem discussion of Section 3.6 — that governs the symmetric (Bayes) error exponent.

The minimizing λ^* and the tilted midpoint distribution. Differentiating $g(\lambda) = \log \sum_x p(x)^\lambda q(x)^{1-\lambda}$ and setting $g'(\lambda^*) = 0$ gives the condition

$$\sum_x Q_{\lambda^*}(x) \log \frac{p(x)}{q(x)} = 0, \qquad Q_{\lambda^*}(x) \propto p(x)^{\lambda^*} q(x)^{1-\lambda^*},$$

i.e. λ^* is chosen so that the tilted distribution Q_{λ^*} is, in a precise sense, "equidistant" from p and q — it is the distribution for which the log-likelihood ratio $\log(p(X)/q(X))$ has mean zero, making it impossible, in expectation, to distinguish which hypothesis generated a sample from Q_{λ^*}. The value $\lambda^* = 1/2$ recovers the *Bhattacharyya* case $C(p,q) \ge -2\log \sum_x \sqrt{p(x)q(x)}$. $\frac{1}{2}$, related to the order-$1/2$ Rényi divergence of Section 1.6, but $\lambda^* = 1/2$ is optimal only when p, q are symmetric in an appropriate sense (e.g. $\lambda^* = 1/2$ for any p, q related by a measure-preserving involution swapping them) — in general λ^* must be found numerically.

Example — Chernoff information for two Bernoulli distributions

$p = \text{Bernoulli}(0.5)$, $q = \text{Bernoulli}(0.7)$ as before. Compute $g(\lambda) = \log_2 \left(p(1)^\lambda q(1)^{1-\lambda} + p(0)^\lambda q(0)^{1-\lambda}\right) = \log_2 \left(0.5^\lambda 0.7^{1-\lambda} + 0.5^\lambda 0.3^{1-\lambda}\right) = \log_2 \left(0.5^\lambda (0.7^{1-\lambda} + 0.3^{1-\lambda})\right)$.
At $\lambda = 0.5$: $0.5^{0.5} = 0.7071$; $0.7^{0.5} = 0.8367$, $0.3^{0.5} = 0.5477$, sum$= 1.3844$. $g(0.5) = \log_2(0.7071 \times 1.3844) = \log_2(0.9790) = -0.0306$. $C \ge -g(0.5) = 0.0306$.
At $\lambda = 0.4$: $0.5^{0.4} = 0.7579$; $0.7^{0.6} = 0.7973$, $0.3^{0.6} = 0.4685$, sum$= 1.2658$. $g(0.4) = \log_2(0.7579 \times 1.2658) = \log_2(0.9593) = -0.0599$. $-g(0.4) = 0.0599$.
At $\lambda = 0.3$: $0.5^{0.3} = 0.8123$; $0.7^{0.7} = 0.7651$, $0.3^{0.7} = 0.4019$, sum$= 1.1670$. $g(0.3) = \log_2(0.8123 \times 1.1670) = \log_2(0.9480) = -0.0771$. $-g(0.3) = 0.0771$.
At $\lambda = 0.2$: $0.5^{0.2} = 0.8706$; $0.7^{0.8} = 0.7336$, $0.3^{0.8} = 0.3437$, sum$= 1.0773$. $g(0.2) = \log_2(0.8706 \times 1.0773) = \log_2(0.9379) = -0.0925$. $-g(0.2) = 0.0925$.
At $\lambda = 0.1$: $0.5^{0.1} = 0.9330$; $0.7^{0.9} = 0.7236$... continuing this scan, $-g(\lambda)$ increases as λ decreases from 0.5 toward 0, suggesting $\lambda^* < 0.2$. Rather than continue the scan by hand, note the qualitative point: because $q(1) = 0.7 > 0.5 = p(1)$, the distribution q is "further" from the uniform-ish p in the direction of favoring outcome 1, and the tilted distributions Q_λ for small λ shift weight toward q; the optimum λ^* for this asymmetric pair is below 0.5. In any case, $C(p,q) \ge 0.0925$ (from $\lambda = 0.2$ alone), giving $P_e^{*(n)} \lesssim 2^{-0.0925n}$ — already a much faster guaranteed decay than the value at $\lambda = 0.5$ would suggest, illustrating why the optimization over λ in the definition of Chernoff information is not a mere technicality: the naive $\lambda = 1/2$ (Bhattacharyya) choice underestimates the true Bayes error exponent by roughly a factor of 3 in this example.

Looking ahead: error exponents in channel coding. The Chernoff–Stein lemma's tilted distributions $Q_\lambda(x) \propto p(x)^\lambda q(x)^{1-\lambda}$ reappear, essentially verbatim, in the random-coding error

exponent for channel coding (Gallager's $E_0(\rho)$ function from the strong-converse discussion of Section 4.5): $E_0(\rho)$ is defined by an optimization over input distributions of a quantity of the form $-\log \sum_y \left(\sum_x p(x) p(y|x)^{1/(1+\rho)} \right)^{1+\rho}$, which is structurally identical to the Chernoff exponent $g(\lambda)$ above with $\lambda = 1/(1+\rho)$ and the roles of "p" and "q" played by the output distributions under two competing codewords. This is not a coincidence: channel decoding *is* a (multi-hypothesis) hypothesis-testing problem — the decoder must decide which of 2^{nR} codewords was sent, based on the channel output — and the pairwise error probability between any two codewords is governed by exactly the Chernoff exponent between their respective output distributions. The method of types, introduced in Section 3.6 to analyze source coding, and developed in this section into the sharp exponents of hypothesis testing, thus turns out to be the single combinatorial tool underlying *both* halves of Shannon's theory — source coding and channel coding — a unification that becomes completely explicit in the error-exponent (rather than merely capacity-achieving) analysis of Volume II.

Part 4

Channel Capacity and Coding Theorems

Discrete memoryless channels, channel capacity, Shannon's channel coding theorem, the BSC and BEC, the Gaussian channel, and the Blahut–Arimoto algorithm.

4.1 Discrete Memoryless Channels

Discrete Memoryless Channel (DMC)

A *discrete memoryless channel* (DMC) is a triple $(\mathcal{X}, p(y|x), \mathcal{Y})$ where $\mathcal{X}$ is the input alphabet, $\mathcal{Y}$ the output alphabet, and $p(y|x)$ the channel transition probabilities. The *memoryless* property means that for any input sequence x^n:

$$p(y^n|x^n) = \prod_{i=1}^{n} p(y_i|x_i).$$

Channel capacity. The *capacity* of a DMC is

$$C = \max_{p(x)} I(X;Y).$$

This maximisation is over all input distributions $p(x)$; the channel $p(y|x)$ is fixed.

Binary Symmetric Channel (BSC). Input and output alphabets $\{0,1\}$; flip probability p: $p(1|0) = p(0|1) = p$, $p(0|0) = p(1|1) = 1-p$. Capacity:

$$C_{\text{BSC}} = 1 - h(p) \text{ bits.}$$

Achieved by $X \sim \text{Uniform}(0,1)$. At $p = 0$: $C = 1$ bits (noiseless). At $p = 1/2$: $C = 0$.

Binary Erasure Channel (BEC). Output $Y \in \{0, 1, ?\}$: with probability ϵ, the bit is erased ($Y = ?$); otherwise received correctly. Capacity:

$$C_{\text{BEC}} = 1 - \epsilon \text{ bits.}$$

Achieved by uniform X. The BEC has capacity strictly higher than the BSC with the same error probability ($1-\epsilon > 1-h(\epsilon)$ for $\epsilon \in (0, 1/2)$) because erasures reveal when the bit was lost, while flips do not.

Example — BSC vs BEC comparison

At noise level $p = \epsilon = 0.1$:

$$C_{\text{BSC}} = 1 - h(0.1) = 1 - 0.469 = 0.531 \text{ bits},$$
$$C_{\text{BEC}} = 1 - 0.1 = 0.9 \text{ bits}.$$

The BEC allows $0.9/0.531 \approx 1.70$ times more reliable information per channel use. The key difference: in the BEC, the receiver knows *when* it has failed; in the BSC, it receives corrupted data without knowing corruption has occurred.

4.2 Shannon's Channel Coding Theorem

Channel Code

An (n, M, δ) *channel code* for channel $p(y|x)$ consists of:

- An encoder $f : [M] \to \mathcal{X}^n$ mapping message index $m \in [M]$ to codeword $x^n(m)$.
- A decoder $g : \mathcal{Y}^n \to [M]$ mapping channel output to message estimate.

The average probability of error is $\bar{P}_e = \frac{1}{M} \sum_m \Pr(g(Y^n) \neq m | X^n = x^n(m))$. The rate is $R = \frac{\log M}{n}$ bits per channel use.

Shannon's Channel Coding Theorem

For a DMC with capacity C:

- **Achievability**: For any $R < C$ and $\epsilon > 0$, there exists n_0 such that for $n \geq n_0$, an $(n, 2^{nR}, \epsilon)$ code exists.
- **Converse**: For any $(n, 2^{nR}, \delta)$ code with $R > C$, $\delta \geq 1 - C/R - O(1/n) > 0$.

Proof of Achievability (Random Coding). *Random codebook generation:* Choose $M = 2^{nR}$ codewords $x^n(1), \ldots, x^n(M)$, each symbol drawn i.i.d. from the capacity-achieving input distribution $p^*(x)$.

Decoder: Joint typical set decoding — declare $\hat{m} = m$ if m is the unique message such that $(x^n(m), Y^n) \in A_\epsilon^{(n)}$ (jointly typical); declare error otherwise.

Error analysis: By symmetry, assume $m = 1$ was sent. Error occurs if (i) $(x^n(1), Y^n)$ is not jointly typical, or (ii) some $m' \neq 1$ has $(x^n(m'), Y^n)$ jointly typical. Event (i): $\Pr(\bar{A}_\epsilon^{(n)}) \to 0$ by the AEP applied to the joint distribution $p(x)p(y|x)$. Event (ii): by the AEP, $\Pr((x^n(m'), Y^n) \in A_\epsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\epsilon)}$ for an independent codeword $x^n(m')$. Summing over $M - 1$ independent codewords: $\Pr(\text{type II error}) \leq (M-1) \cdot 2^{-n(I(X;Y)-3\epsilon)} \leq 2^{n(R-I(X;Y)+3\epsilon)} \to 0$ for $R < I(X;Y) - 3\epsilon = C - 3\epsilon$. Since the expected error (over random codes) is small, there exists a deterministic code with the same error bound.

Proof of Converse (via Fano's Inequality). Let M be uniformly distributed on $[2^{nR}]$. The channel output Y^n depends on M only through X^n. By Fano's inequality:

$$H(M|Y^n) \leq 1 + P_e \cdot nR.$$

Also $H(M|Y^n) = H(M) - I(M;Y^n) = nR - I(M;Y^n)$. By the chain rule and the DPI:

$$I(M;Y^n) = I(X^n;Y^n) = \sum_{i=1}^{n} I(X_i;Y^n|X^{i-1}) \leq \sum_i I(X_i;Y_i) \leq nC.$$

(The last inequality uses $I(X_i;Y_i) \leq C$ for each i.) Therefore: $nR - nC \leq H(M|Y^n) \leq 1 + P_e \cdot nR$, giving $P_e \geq (R-C)/R - 1/(nR) \to (R-C)/R > 0$.

Example — Capacity and coding gap for the BSC at $p = 0.1$

$C = 0.531$ bits/channel use. To transmit a message of $k = 100$ bits reliably using n channel uses, we need $R = k/n < C$, so $n > k/C = 100/0.531 \approx 188$ channel uses. The *coding overhead* is $n - k = 88$ redundant bits. A rate-1/2 code ($n = 200$, $k = 100$) operates at $R = 0.5 < 0.531$ and can (in principle) achieve arbitrarily small

error probability for large block lengths. Without coding ($n = k = 100$, rate 1), each bit is received correctly with probability 0.9; by independence, $\Pr(\text{all 100 correct}) = 0.9^{100} \approx 2.66 \times 10^{-5}$ — extremely unreliable. Coding transforms the channel into a virtually error-free link.

4.3 The Gaussian Channel and Water-Filling

Additive White Gaussian Noise Channel

The *AWGN channel* with power constraint P and noise variance N is:

$$Y_i = X_i + Z_i, \quad Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, N), \quad \mathbb{E}[X_i^2] \le P.$$

Its capacity is

$$C = \frac{1}{2}\log_2\left(1 + \frac{P}{N}\right) \text{ bits/channel use.}$$

Proof. *Upper bound*: $I(X;Y) = h(Y) - h(Y|X) = h(Y) - h(Z) \le \frac{1}{2}\log(2\pi e(P+N)) - \frac{1}{2}\log(2\pi eN) = \frac{1}{2}\log(1 + P/N)$, using $h(Y) \le \frac{1}{2}\log(2\pi e\,\mathrm{Var}[Y])$ and $\mathrm{Var}[Y] = P + N$.

Lower bound (achievability): Choose $X \sim \mathcal{N}(0, P)$, making $Y \sim \mathcal{N}(0, P+N)$. Then $h(Y) = \frac{1}{2}\log(2\pi e(P+N))$ and $h(Z) = \frac{1}{2}\log(2\pi eN)$, giving $I(X;Y) = \frac{1}{2}\log(1 + P/N)$.

Band-limited channel capacity (Shannon–Hartley theorem). For a band-limited channel of bandwidth W Hz, signal power P, and noise power spectral density $N_0/2$ (so noise power in W Hz is $N = N_0 W$):

$$C = W\log_2\left(1 + \frac{P}{N_0 W}\right) \text{ bits/second.}$$

Example — Shannon capacity of telephone line

Bandwidth $W = 3400$ Hz, SNR $= P/N = 1000$ (30 dB). Noise power $N = N_0 W$. $C = 3400\log_2(1001) \approx 3400 \times 9.97 = 33{,}900$ bits/second ≈ 33.9 kbps. A 56 kbps modem operates at SNR higher than assumed here (using line coding on POTS), while fibre optics use W in the terahertz range.

Parallel Gaussian channels and water-filling. For K independent parallel AWGN channels with noise variances $N_1, \dots, N_K$ and total power constraint $\sum_k P_k \le P$, the total capacity is

$$C = \sum_{k=1}^{K} \frac{1}{2}\log_2\left(1 + \frac{P_k^*}{N_k}\right),$$

where $P_k^* = (\mu - N_k)^+$ and μ is chosen such that $\sum_k P_k^* = P$ (water-filling).

Water-filling derivation. Maximise $\sum_k \frac{1}{2}\log(1 + P_k/N_k)$ subject to $\sum_k P_k = P$, $P_k \ge 0$. Lagrangian:

$$\mathcal{L} = \sum_k \frac{1}{2}\log(1 + P_k/N_k) - \lambda\left(\sum_k P_k - P\right).$$

KKT conditions: $\frac{1}{2}\cdot\frac{1}{N_k + P_k} = \lambda$, so $P_k + N_k = 1/(2\lambda) = \mu$, giving $P_k^* = (\mu - N_k)^+$. The water level μ is a "water" poured over the "ground" of noise levels N_k; channels with $N_k > \mu$ receive no power.

Example — Water-filling over 3 parallel channels

$N_1 = 1$, $N_2 = 2$, $N_3 = 5$, total power $P = 6$. Water level μ: try $\mu = 4$: $P_1^* = 3$, $P_2^* = 2$, $P_3^* = 0$. Check: $3 + 2 + 0 = 5 \neq 6$. Try $\mu = 4.5$: $P_1^* = 3.5$, $P_2^* = 2.5$, $P_3^* = 0$. Check: $3.5 + 2.5 = 6$. ✓. Channel 3 ($N_3 = 5 > 4.5$) receives no power. Capacity: $C = \frac{1}{2}\log_2(1 + 3.5/1) + \frac{1}{2}\log_2(1 + 2.5/2) = \frac{1}{2}\log_2(4.5) + \frac{1}{2}\log_2(2.25) = \frac{1}{2}(2.170) + \frac{1}{2}(1.170) = 1.085 + 0.585 = 1.670$ bits/use.

4.4 The Blahut–Arimoto Algorithm

Blahut–Arimoto Algorithm for Channel Capacity

Given a DMC $p(y|x)$, the Blahut–Arimoto algorithm computes $C = \max_{p(x)} I(X;Y)$ iteratively:

1. Initialise $q^{(0)}(x)$ uniformly on $\mathcal{X}$.
2. For $t = 0, 1, 2, \ldots$: compute

$$c^{(t)}(x) = \exp\left(\sum_y p(y|x)\log\frac{p(y|x)}{\sum_{x'} q^{(t)}(x')p(y|x')}\right),$$

   then update $q^{(t+1)}(x) = c^{(t)}(x)/\sum_{x'} c^{(t)}(x')$.
3. Stop when $\max_x \log c^{(t)}(x) - \log\left(\sum_{x'} c^{(t)}(x')\right) < \epsilon$.

The upper bound $U^{(t)} = \log\sum_{x'} c^{(t)}(x')$ and lower bound $L^{(t)} = \sum_x q^{(t)}(x)\log c^{(t)}(x)$ satisfy $L^{(t)} \leq C \leq U^{(t)}$, and $U^{(t)} - L^{(t)} \to 0$.

Convergence. The algorithm converges to the global maximum because $I(X;Y)$ is concave in $p(x)$: the algorithm is a coordinate ascent in the form of an alternating optimisation on the dual representation of capacity. The convergence rate is linear (geometric), with rate determined by the second eigenvalue of the channel matrix.

Example — Blahut–Arimoto for a 3×3 channel

Channel matrix $p(y|x)$ with $|\mathcal{X}| = |\mathcal{Y}| = 3$:

$$P = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.1 & 0.7 & 0.2 \\ 0.2 & 0.1 & 0.7 \end{pmatrix}.$$

Starting from $q^{(0)} = (1/3, 1/3, 1/3)$, after 20 iterations the algorithm converges to $q^* \approx (0.342, 0.329, 0.329)$ with $C \approx 0.521$ bits. The uniform input is nearly optimal, reflecting the near-symmetry of the channel. Each column of P concentrates probability on the diagonal, consistent with a capacity well below $\log_2 3 = 1.585$ bits.

4.5 Strong Converse and the Channel Dispersion

The strong converse. Shannon's converse shows $P_e \to 1$ as $n \to \infty$ for $R > C$. A *strong converse* shows this convergence is exponentially fast: $P_e \geq 1 - e^{-n(R-C)^2/(2\sigma^2)+o(n)}$ where σ^2 is the channel dispersion. The BSC and BEC both satisfy the strong converse.

Channel Dispersion

The *channel dispersion* V is the variance of the information density:

$$V = \mathrm{Var}_{p^*(x)}\left[\log \frac{p(Y|X)}{p(Y)}\right],$$

evaluated at the capacity-achieving input distribution $p^*(x)$.

Finite-blocklength channel coding (Polyanskiy–Poor–Verdú, 2010). The maximum rate achievable at block length n with error probability ϵ satisfies

$$R^*(n,\epsilon) = C - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon) + O\left(\frac{\log n}{n}\right),$$

where $Q^{-1}(\epsilon)$ is the inverse Gaussian Q-function. This second-order result shows:

- At $\epsilon = 0.01$: the overhead is $2.326\sqrt{V/n}$ bits/channel use.
- The dispersion V determines how quickly capacity is approached as n increases.
- For the BSC with $p = 0.1$: $V = p(1-p)(\log[(1-p)/p])^2 = 0.09 \times (2.807)^2 = 0.709$ bits^2.

Example — Finite-blocklength rate for BSC

BSC with $p = 0.1$, $C = 0.531$ bits, $V = 0.709$ bits^2. Target $\epsilon = 10^{-3}$, $Q^{-1}(10^{-3}) \approx 3.09$.
At $n = 100$: $R^*(100, 10^{-3}) \approx 0.531 - 3.09\sqrt{0.709/100} = 0.531 - 3.09 \times 0.0842 = 0.531 - 0.260 = 0.271$ bits. Only about half the asymptotic capacity is achievable.
At $n = 1000$: $R^* \approx 0.531 - 3.09\sqrt{0.709/1000} = 0.531 - 3.09 \times 0.0266 = 0.531 - 0.082 = 0.449$ bits. Now 85% of capacity is achievable.
At $n = 10{,}000$: $R^* \approx 0.531 - 0.026 = 0.505$ bits (95% of C). The $\sqrt{V/n}$ penalty highlights why modern codes (5G, satellite) use blocks of at least $n = 1000$–$10{,}000$ symbols.

Meta-converse. The Polyanskiy–Poor–Verdú framework provides a unified "meta-converse" that recovers all classical converses (weak, strong, sphere packing) as special cases, using a hypothesis-testing formulation. The *$\kappa\beta$ bound* states:

$$M^* \leq \frac{1}{\beta_{1-\epsilon}(p_0, p_1)},$$

where $\beta_\alpha(p_0, p_1)$ is the minimum type-II error probability at type-I error $\leq \alpha$ in the hypothesis test $H_0 : p_0$ vs $H_1 : p_1$.

Summary: the fundamental limits of information theory. This volume established the three foundational results:

1. **Entropy** $H(X)$ is the irreducible uncertainty in a discrete random variable; its properties (non-negativity, maximality, chain rule, concavity) are all consequences of Jensen's inequality and the log-sum inequality.
2. **Shannon's source coding theorem**: the entropy rate $\mathcal{H}$ is the minimum compression rate for a stationary ergodic source; the AEP is the key analytical tool.
3. **Shannon's channel coding theorem**: the channel capacity $C = \max_{p(x)} I(X;Y)$ is the maximum reliable communication rate; random coding achieves it, and Fano's inequality proves the converse.

Volume II develops the constructive side: how to *build* codes that approach these limits.

4.6 Channels with Memory, Feedback, and the Information Spectrum Method

The channel coding theorem of Section 4.2 was proved for *discrete memoryless* channels: each output Y_i depends only on the corresponding input X_i, and successive uses of the channel are statistically independent given the inputs. Real channels — wireless links with fading correlated across time, magnetic recording channels with inter-symbol interference, queueing systems with bursty packet loss — violate this assumption. This section extends the capacity formula to channels with memory, examines whether feedback from receiver to transmitter can increase capacity, and introduces the *information spectrum method*, a general framework (due to Verdú and Han) that characterizes the capacity of *arbitrary* channels — including those that are non-stationary or non-ergodic, for which the simple formula $C = \max_p I(X;Y)$ does not even make sense as stated — purely in terms of the asymptotic behavior of a sequence of information densities. This generality will matter in Volume II, where real coding schemes must be designed for channels whose statistics are only approximately known or are slowly time-varying.

Capacity of a Channel with Memory

For a channel $\{p(y^n|x^n)\}_{n\geq 1}$ (a sequence of conditional distributions, not necessarily arising from single-letter $p(y|x)$ applied independently), the capacity is

$$C = \lim_{n\to\infty} \frac{1}{n} \max_{p(x^n)} I(X^n;Y^n),$$

provided the limit exists. If the channel is *stationary and has finite memory* (the output Y_i depends on $(X_i, \dots, X_{i-k})$ for some fixed k, and the dependence is time-invariant), the limit exists and equals

$$C = \lim_{n\to\infty} \left[\max_{p(x^n)} I(X^n;Y^n) - \max_{p(x^{n-1})} I(X^{n-1};Y^{n-1})\right] = \sup_n \max_{p(x_n|x^{n-1})} \frac{1}{n} I(X^n;Y^n)$$

(the second equality, a Cesàro-mean identity, holds because the sequence $a_n = \max_{p(x^n)} I(X^n;Y^n)$ is superadditive: $a_{n+m} \geq a_n + a_m$, so a_n/n converges to $\sup_n a_n/n$ by Fekete's lemma — the superadditive counterpart of the subadditive Fekete's lemma used for entropy rate in Section 1.5).

Why memory complicates the maximization. For a memoryless channel, $I(X^n;Y^n) \leq \sum_i I(X_i;Y_i)$ (by the chain rule and the fact that Y_i depends only on X_i, so conditioning on other X_j, Y_j cannot increase $I(X_i;Y_i)$), with equality for product input distributions — this is exactly why the single-letter formula $C = \max_{p(x)} I(X;Y)$ suffices in Section 4.2. With memory, Y_i depends on past inputs, so the chain-rule decomposition $I(X^n;Y^n) = \sum_i I(X_i;Y^n|X^{i-1})$ does not collapse to a sum of single-letter terms, and the maximizing input distribution $p(x^n)$ is generally *not* a product distribution — the optimal code must exploit correlations across time to "pre-compensate" for the channel's memory (a phenomenon familiar from water-filling over a frequency-selective Gaussian channel with memory, which is exactly the continuous-alphabet instance of this principle, generalizing the parallel-Gaussian water-filling of Section 4.3 from a finite set of parallel sub-channels to a continuum indexed by frequency via the Fourier transform of the channel's memory kernel).

Does feedback help? Shannon's feedback theorem for the DMC. Suppose the encoder, at each time i, observes the entire past output sequence Y^{i-1} before choosing X_i (noiseless

feedback). Does this increase capacity?

Shannon's Feedback Capacity Theorem (DMC)

For a discrete memoryless channel, feedback does not increase capacity:

$$C_{\text{FB}} = C = \max_{p(x)} I(X;Y).$$

Proof. The converse direction ($C_{\text{FB}} \geq C$ is trivial, since any non-feedback code is in particular a feedback code that ignores the feedback) requires only the achievability direction $C_{\text{FB}} \leq C$. By Fano's inequality (Section 4.2) applied to the feedback system, $nR \leq I(X^n;Y^n) + n\epsilon_n$ for a rate-R code with vanishing error probability, where now X^n is generated causally with X_i a function of the message W and Y^{i-1}. The key step is bounding $I(X^n;Y^n)$ *despite* the feedback-induced dependence between X_i and past outputs:

$$I(X^n;Y^n) = \sum_{i=1}^{n} I(X_i;Y_i|Y^{i-1}) \leq \sum_{i=1}^{n} I(X_i, Y^{i-1};Y_i) = \sum_{i=1}^{n} \big[I(Y^{i-1};Y_i) + I(X_i;Y_i|Y^{i-1})\big],$$

which seems circular; the correct route (due to Shannon's original 1956 argument) instead bounds

$$I(X^n;Y^n) = H(Y^n) - H(Y^n|X^n) = \sum_{i=1}^{n} \big[H(Y_i|Y^{i-1}) - H(Y_i|X^n, Y^{i-1})\big].$$

Since the channel is memoryless, $H(Y_i|X^n, Y^{i-1}) = H(Y_i|X_i)$ (the output at time i depends only on the input at time i, regardless of what else — past outputs, future inputs — is known). And $H(Y_i|Y^{i-1}) \leq H(Y_i)$ always. Therefore

$$I(X^n;Y^n) \leq \sum_{i=1}^{n} \big[H(Y_i) - H(Y_i|X_i)\big] = \sum_{i=1}^{n} I(X_i;Y_i) \leq nC,$$

where the last step bounds each term by the single-letter capacity $C = \max_p I(X;Y)$. Combining with Fano's inequality, $nR \leq nC + n\epsilon_n$, so $R \leq C + \epsilon_n \to C$.

Interpretation. The proof identifies precisely *why* feedback fails to help for the DMC: the memorylessness of the channel means $H(Y_i|X^n, Y^{i-1}) = H(Y_i|X_i)$ — the conditional output entropy at time i cannot be reduced by feedback information, because the channel "forgets" everything except the current input. Feedback *does* help, dramatically, in two related but distinct senses that the theorem does not rule out: (1) it can simplify the *encoder* enormously (the Schalkwijk–Kailath scheme achieves the Gaussian channel capacity with a trivial linear encoder and doubly-exponentially small error probability, versus the much more complex random-coding constructions needed without feedback); and (2) for channels *with memory*, feedback *can* increase capacity, because $H(Y_i|X^n, Y^{i-1})$ need not reduce to $H(Y_i|X_i)$ when the channel state at time i depends on the history — the encoder, knowing past outputs, can infer the channel's current state and adapt, something it cannot do from X^n alone if the state process is not a deterministic function of past inputs (e.g. if the channel has an exogenous Markov state, as in the Gilbert–Elliott model below).

Example — A two-state channel with memory: the Gilbert–Elliott model

Consider a binary symmetric channel whose crossover probability switches between two values according to a hidden Markov chain — a simple model of a wireless link alternating between "good" and "bad" states. Let the state $S_i \in \{G, B\}$ follow a sta-

tionary Markov chain with $\Pr(S_{i+1} = B|S_i = G) = g$, $\Pr(S_{i+1} = G|S_i = B) = b$, stationary probabilities $\pi_G = \frac{b}{g+b}$, $\pi_B = \frac{g}{g+b}$. Given $S_i = G$, the channel is a BSC with crossover ϵ_G; given $S_i = B$, crossover $\epsilon_B > \epsilon_G$.
Numerical instance: $g = 0.1$, $b = 0.3$ (so $\pi_G = 0.75$, $\pi_B = 0.25$ — the channel is in the good state 75% of the time), $\epsilon_G = 0.01$, $\epsilon_B = 0.3$.
Capacity if the decoder knows the state sequence S^n (genie-aided): $C_{\text{genie}} = \pi_G(1 - h(\epsilon_G)) + \pi_B(1 - h(\epsilon_B))$. $h(0.01) = 0.0808$, $h(0.3) = 0.8813$.

$C_{\text{genie}} = 0.75(1-0.0808)+0.25(1-0.8813) = 0.75(0.9192)+0.25(0.1187) = 0.6894+0.0297 = 0.7191$ bits.

Capacity if the state is hidden (the realistic case): $C_{\text{hidden}} < C_{\text{genie}}$ strictly, because the receiver must additionally estimate the state from the noisy output sequence, and this estimation is imperfect. C_{hidden} does not have a closed form in general — it requires the limit $\lim_n \frac{1}{n} I(X^n; Y^n)$ for the maximizing input process, which for a hidden-Markov channel is typically evaluated numerically via the Arimoto–Blahut algorithm adapted to the state-augmented channel, or via simulation-based estimators of the entropy rate $H(Y^n|X^n)$ using the forward algorithm for hidden Markov models. What *can* be said in closed form is the genie-aided bound above, which is an upper bound on C_{hidden} (extra side information cannot decrease capacity), and a lower bound via $C_{\text{hidden}} \geq 1 - [\pi_G h(\epsilon_G) + \pi_B h(\epsilon_B)]$ — wait, this naive average-crossover bound is generally *not* valid as a lower bound for capacity (mutual information is not linear in the channel parameters), but the time-averaged crossover $\bar{\epsilon} = \pi_G \epsilon_G + \pi_B \epsilon_B = 0.75(0.01) + 0.25(0.3) = 0.0825$ gives, via the memoryless-BSC formula evaluated at $\bar{\epsilon}$, $1 - h(0.0825) = 1 - 0.403 = 0.597$, a commonly-used (but not rigorously justified) engineering estimate that brackets the true C_{hidden} between roughly 0.597 and 0.7191 bits per channel use.
Effect of feedback. If the encoder receives noiseless feedback of Y^{i-1}, it can — in principle — run the same state-estimation algorithm as the receiver and adapt its input distribution (e.g. transmitting at higher rate when the estimated state is G). This can strictly increase achievable rates relative to a single fixed input distribution used for both states, illustrating point (2) above: feedback helps for channels with memory in a way it provably cannot for the memoryless channel of Shannon's feedback theorem.

The information spectrum method. For genuinely general channels — not necessarily stationary, ergodic, or even possessing a well-defined single-letter capacity formula — Verdú and Han (1994) showed that the capacity is still given by a clean formula, provided one replaces the limit in the definition of capacity by a more general statistical quantity: the *liminf in probability*.

Information Spectrum / Liminf in Probability

For a sequence of real-valued random variables $\{Z_n\}$, the *liminf in probability* is

$$\text{p-}\liminf_{n\to\infty} Z_n = \sup\left\{\alpha \in \mathbb{R} : \lim_{n\to\infty} \Pr(Z_n < \alpha) = 0\right\}.$$

Equivalently, it is the largest α such that $Z_n \geq \alpha - \delta$ with probability $\to 1$ for every $\delta > 0$.

Verdú–Han General Channel Capacity Formula

For an arbitrary channel $\{p(y^n|x^n)\}$ (possibly non-stationary, non-ergodic), the capacity is

$$C = \sup_{\{p(x^n)\}} \operatorname{p-lim\,inf}_{n\to\infty} \frac{1}{n} \log \frac{p(Y^n|X^n)}{p(Y^n)},$$

where the supremum is over sequences of input distributions $p(x^n)$, and $\frac{1}{n} \log \frac{p(Y^n|X^n)}{p(Y^n)}$ is the *normalized information density* — a random variable (a function of the random pair (X^n, Y^n)), not a deterministic quantity.

Recovering the classical formula. For a stationary memoryless channel with a fixed (non-sequence-dependent) input distribution $p(x)$ used i.i.d. across the block, the normalized information density $\frac{1}{n} \log \frac{p(Y^n|X^n)}{p(Y^n)} = \frac{1}{n} \sum_i \log \frac{p(Y_i|X_i)}{p(Y_i)}$ is, by the law of large numbers, an average of i.i.d. terms converging *in probability* (indeed almost surely) to its expectation $I(X;Y)$. For a quantity converging in probability to a constant c, the liminf in probability equals c exactly (the defining supremum is achieved at $\alpha = c$, since $\Pr(Z_n < \alpha) \to 0$ for $\alpha \le c$ and $\Pr(Z_n < \alpha) \to 1$ for $\alpha > c$). Hence the general formula reduces to $C = \sup_p I(X;Y) = \max_p I(X;Y)$, recovering Section 4.2 exactly — the information spectrum formula is a strict generalization, not a different theory.

Why the liminf-in-probability, not the expectation? For a non-ergodic channel, the normalized information density may converge (in probability) to a *random* limit — e.g. the channel might be, with probability $1/2$ each, one of two fixed DMCs with different capacities $C_1 < C_2$ (chosen once, at time 0, and then fixed for the whole transmission — a "mixed channel"). The *expected* information density would average to $\frac{1}{2}(C_1 + C_2)$, but no code can reliably achieve a rate above C_1: with probability $1/2$ the channel is the worse one, and any rate above C_1 fails on that half of the sample space, so the error probability cannot vanish. The liminf in probability correctly identifies $C = C_1$ (the α such that $\Pr(Z_n < \alpha) \to 0$ requires $\alpha \le C_1$, since on the bad-channel half of the sample space $Z_n \to C_1 < C_2$), whereas the naive expectation $\frac{1}{2}(C_1 + C_2)$ would be an overly optimistic, unachievable answer. This distinction — between "capacity" as an *almost-sure* guarantee versus an *average-case* quantity — is exactly analogous to the distinction between min-entropy and Shannon entropy from Section 1.6: worst-case (or, here, "almost-sure-case") operational guarantees require the more conservative, order-statistics-like functional, not the mean.

The general converse and achievability via information spectrum. The full Verdú–Han theory proves both directions of the formula via non-asymptotic bounds that generalize Fano's inequality and the random-coding bound:

Converse. For *any* code of rate R and block length n, the error probability satisfies $P_e^{(n)} \ge \Pr\left(\frac{1}{n} \log \frac{p(Y^n|X^n)}{p(Y^n)} \le R - \gamma\right) - 2^{-n\gamma}$ for any $\gamma > 0$ — a direct, non-asymptotic relationship between error probability and the information density, with no appeal to typicality or ergodicity.

Achievability. A random-coding argument (codewords drawn i.i.d. from the optimizing input process, exactly as in Section 4.2's proof) shows $P_e^{(n)} \le \Pr\left(\frac{1}{n} \log \frac{p(Y^n|X^n)}{p(Y^n)} \le R + \gamma\right) + 2^{-n\gamma}$, again non-asymptotic and requiring no structural assumptions on the channel beyond the existence of the relevant conditional and marginal distributions $p(y^n|x^n)$, $p(y^n)$.

Both bounds, in the limit $n \to \infty$, pin down C as the liminf in probability of the information density — converse and achievability meet exactly at the threshold defined by the information-spectrum formula, with no gap, for *any* channel.

Looking ahead. The information spectrum method will be essential in Volume II for analyzing the *finite-blocklength* performance of practical codes — the second-order (dispersion) terms in the asymptotic expansion $\log M^*(n,\epsilon) = nC - \sqrt{nV}\,Q^{-1}(\epsilon) + O(\log n)$ mentioned in the strong-converse discussion of Section 4.5 are themselves derived by applying a Berry–Esseen-type refinement to the information-spectrum formula, replacing the liminf in probability by a precise Gaussian approximation to the distribution of the normalized information density at finite n. The channel-dispersion quantity V is, in this language, simply the variance of the (single-letter) information density $\log \frac{p(Y|X)}{p(Y)}$ under the capacity-achieving input distribution — connecting the asymptotic spectrum-theoretic viewpoint of this section to the concrete, computable, second-order channel characterization that governs code design at realistic block lengths.

4.7 Joint Source-Channel Coding and the Separation Theorem

Parts 3 and 4 developed, in parallel, the two halves of Shannon's theory: source coding (compressing a source to its entropy rate $\mathcal{H}$) and channel coding (transmitting information reliably at rates up to capacity C). This closing section of Volume I asks the question that justifies treating these as separate problems at all: if the ultimate goal is to convey samples from a source $\{U_i\}$, with entropy rate $\mathcal{H}$, reliably across a noisy channel of capacity C, is it ever beneficial to design a single *joint* encoder/decoder pair tailored to both the source statistics and the channel statistics simultaneously, rather than simply cascading an optimal source code with an optimal channel code? Shannon's separation theorem answers this question — asymptotically, and for the point-to-point setting of this volume — with a resounding "no," but the precise conditions under which separation is optimal, and the ways it fails outside those conditions, are themselves deeply informative, and set the stage for the network information theory previewed in Appendix E and developed properly in later volumes.

Joint Source-Channel Coding Theorem (Separation Theorem)

Let $\{U_i\}$ be a stationary ergodic source with entropy rate $\mathcal{H}$ (Section 1.5), and let the channel have capacity C (Section 4.2). A sequence of source samples U^n can be encoded into n channel uses and decoded with vanishing block error probability ($\Pr(\hat{U}^n \neq U^n) \to 0$) if

$$\mathcal{H} < C,$$

and cannot be so encoded (i.e. $\Pr(\hat{U}^n \neq U^n) \not\to 0$) if $\mathcal{H} > C$. Moreover, this threshold is achieved by *separate* source and channel coding: compress U^n to $\approx n\mathcal{H}$ bits using an optimal source code (Part 3), then transmit those bits using an optimal channel code at any rate R with $\mathcal{H} < R < C$ (Part 4).

Achievability via separation. Fix any rate $R \in (\mathcal{H}, C)$. By the source coding theorem (Section 3.3), U^n can be losslessly compressed (with vanishing error probability — strictly, vanishing probability that the source sequence falls outside the typical set being losslessly encoded) into a block of $\approx nR$ bits using a code whose rate is $R > \mathcal{H}$ (any rate above the entropy rate suffices for the existence of such a code, by the converse direction's slack). By the channel coding theorem (Section 4.2), since $R < C$, these nR bits can be transmitted across n channel uses with error probability $\to 0$. The composition — source encoder, channel encoder, channel, channel decoder, source decoder — therefore has overall error probability $\to 0$ by the union bound over the (vanishing) error events of each stage.

The converse: why $\mathcal{H} > C$ is fatal, via data processing. Suppose $\mathcal{H} > C$ and consider *any* encoder $f_n : \mathcal{U}^n \to \mathcal{X}^n$ and decoder $g_n : \mathcal{Y}^n \to \mathcal{U}^n$ (not assumed to be of the separated

form). The chain $U^n \to X^n = f_n(U^n) \to Y^n \to \hat{U}^n = g_n(Y^n)$ is a Markov chain, so by the data-processing inequality (Section 2.3),

$$I(U^n;\hat{U}^n) \le I(X^n;Y^n) \le nC$$

(the second inequality because $I(X^n;Y^n) \le n \max_p I(X;Y) = nC$ for a memoryless channel, by the same chain-rule argument used in Section 4.6 to bound capacity with memory — here there is no memory, so the bound is simply n times the single-letter capacity). On the other hand, if $\Pr(\hat{U}^n = U^n) \to 1$, Fano's inequality (Section 4.2) gives $H(U^n|\hat{U}^n) \le 1 + \Pr(\hat{U}^n \neq U^n)\log|\mathcal{U}|^n = o(n)$, so

$$I(U^n;\hat{U}^n) = H(U^n) - H(U^n|\hat{U}^n) \ge H(U^n) - o(n) = n\mathcal{H} - o(n)$$

(using $H(U^n)/n \to \mathcal{H}$, the definition of entropy rate). Combining, $n\mathcal{H} - o(n) \le nC$, i.e. $\mathcal{H} \le C + o(1)$. If $\mathcal{H} > C$ strictly, this is a contradiction for n large enough, so $\Pr(\hat{U}^n = U^n) \not\to 1$.

The remarkable feature of this argument: no separation was assumed. The converse proof did not assume that the encoder f_n first compresses U^n and then channel-encodes the result — f_n could be an arbitrary, deeply "joint" mapping from source sequences directly to channel inputs, exploiting any conceivable correlation structure. The data-processing inequality bounds $I(U^n;\hat{U}^n)$ by $I(X^n;Y^n) \le nC$ *regardless* of how X^n depends on U^n. This is the precise sense in which separation is *lossless* asymptotically: no joint scheme, however clever, can beat the threshold $\mathcal{H} < C$ that the simple cascade of separately-optimal source and channel codes already achieves. The entire benefit that joint design *could* in principle offer — exploiting source structure to shape the channel-input distribution, or exploiting channel structure to tolerate certain source errors more gracefully than others — evaporates in the limit $n \to \infty$ for point-to-point, single-source, single-channel transmission with the average-error and entropy-rate criteria used here.

Example — Separation in action: a biased binary source over a BSC

Let $\{U_i\}$ be i.i.d. Bernoulli(0.1) (so $\mathcal{H} = h(0.1) = 0.469$ bits/symbol), and let the channel be a BSC with crossover $\epsilon = 0.05$, giving capacity $C = 1 - h(0.05) = 1 - 0.286 = 0.714$ bits/use.

Separation scheme. Since $\mathcal{H} = 0.469 < C = 0.714$, reliable transmission is possible. Concretely: for every 1000 source symbols, an optimal source code (e.g. a Huffman or arithmetic code approaching the entropy bound, Section 3.3) produces ≈ 469 bits. These 469 bits are then channel-coded at rate R with $0.469 < R < 0.714$ — say $R = 0.6$ — requiring $469/0.6 \approx 782$ channel uses, well within the 1000 available if the channel can be used, e.g., 1000 times per 1000 source symbols (a 1-to-1 source-symbol-to-channel-use ratio is the natural normalization when both are specified "per unit time"; the slack $1000 - 782 = 218$ channel uses represents unused capacity, which could instead be used to lower R further, trading off decoding complexity and latency against the (asymptotically irrelevant, but practically real) finite-blocklength error probability).

What if the source were less compressible? If instead $U_i \sim$ Bernoulli(0.4), $\mathcal{H} = h(0.4) = 0.971$ bits/symbol $> C = 0.714$. By the converse direction, *no* scheme — joint or separated — can achieve $\Pr(\hat{U}^n = U^n) \to 1$. The source simply contains more information per symbol than the channel can carry per use, and this is a hard, scheme-independent limit. (Lossy joint source-channel coding — allowing $\hat{U}^n \neq U^n$ with bounded average distortion, governed by the rate-distortion function $R(D)$ rather than

$\mathcal{H}$, briefly touched in the Open Problems appendix — relaxes this limit by accepting a controlled amount of reconstruction error, but that is a different (and, for this volume, out-of-scope) problem.)

Where separation fails: a preview of network information theory. The separation theorem is a statement about *point-to-point* communication: one source, one encoder, one channel, one decoder. The moment the topology becomes more complex — multiple sources sharing a channel (multiple access), one source broadcast to multiple receivers with different channel qualities, or a source observed with side information available only at the decoder (Wyner–Ziv-type problems) — separation generically *fails*: jointly designed source-channel codes can strictly outperform any cascade of separately-optimal source and channel codes. The canonical example is broadcasting a Gaussian source over a Gaussian broadcast channel to two receivers with different SNRs: a separated scheme must pick a single compression rate for the source, achievable by the weaker receiver, and the stronger receiver gains nothing from its better channel; a joint scheme (e.g. superposition coding combined with successive-refinement source coding) lets the stronger receiver obtain a strictly better reconstruction by exploiting the "extra" channel capacity it has relative to the weaker receiver — extra capacity that a separated scheme cannot route back into improved source fidelity because the source code was designed independently of the channel. Appendix E's preview of network information theory touches on the multiple-access and broadcast capacity *regions*; the failure of separation in these settings is one of the deepest reasons multi-terminal information theory remains, in many cases, open even seventy years after Shannon's original work — there is no general analogue of the separation theorem to fall back on, and capacity regions must be determined (or bounded) on a case-by-case basis.

Mismatch: universal coding meets channel coding. A second, more subtle failure mode of separation arises even in the point-to-point setting when the source statistics are not known exactly at design time. Section 3.6 showed that universal source codes (via the method of types, or Lempel–Ziv) achieve the entropy rate $\mathcal{H}(p)$ for the *true* source distribution p with vanishing redundancy, *without knowing* p. An analogous question for channels — is there a universal channel *code* that achieves capacity $C(p_{Y|X})$ for an unknown channel law, without knowing $p_{Y|X}$? — has a more delicate answer: for a compound channel (the true channel is one of a known set $\{p^{(\theta)}_{Y|X}\}_{\theta\in\Theta}$, Θ possibly infinite), the compound channel capacity is $\min_\theta C(p^{(\theta)}_{Y|X})$ — the capacity of the *worst* channel in the set — rather than the capacity of the actual channel, which could be much larger. This is the channel-coding analogue of the min-entropy H_∞ from Section 1.6 versus the Shannon entropy H: universal *source* coding (an average-case, entropy-rate guarantee) and universal *channel* coding under uncertainty (a worst-case, min-capacity guarantee) are governed by different order statistics of the underlying parameter family, and a separated scheme built from an entropy-rate-optimal source code and a min-capacity-optimal channel code can be strictly suboptimal compared to a joint scheme that adapts its compression rate to the (a priori unknown, but possibly estimable online) channel state — another crack in the separation theorem's idealized assumptions, and a recurring theme in the design of real adaptive communication systems.

Closing the loop: the architecture of Volume I. With this section, the four parts of Volume I form a complete, closed picture of point-to-point communication under a single, unifying quantity — entropy, in its many guises. Part 1 defined entropy and its generalizations (Section 1.6); Part 2 showed that every notion of "distance between distributions" relevant to information theory is an instance of f-divergence (Section 2.5) and that every "natural" parametric family arises from maximizing entropy under constraints, with KL divergence as the resulting Bregman geometry (Section 2.6); Part 3 showed that the entropy rate is

both the compression limit (Section 3.3) and — via the method of types (Section 3.6) and its hypothesis-testing consequences (Section 3.7) — the exponential rate governing every large-deviation question about i.i.d. data; and Part 4 showed that channel capacity, the dual quantity to entropy rate, governs reliable transmission even in the presence of memory and feedback (Section 4.6), with the separation theorem of this section showing that the two quantities $\mathcal{H}$ and C, derived by entirely independent arguments in Parts 3 and 4, are precisely the two numbers that need to be compared to answer the most basic question communication theory can ask: can this source be sent over this channel? Volume II takes up the constructive question — *how* — building the actual codes (linear codes, LDPC codes, turbo codes, arithmetic and universal codes) that approach the limits $\mathcal{H}$ and C characterized here, while Volume III revisits the convex-duality and Bregman-divergence structures of Section 2.6 as the foundation of a full Riemannian geometry on the space of probability distributions.

Numerical Tour: Part 1 — From Distributions to Entropy Rates

This tour computes every entropy quantity introduced in Part 1 for a single running example, then extends the calculation to continuous and process-level entropy.

The running distribution. Let X take values in $\{1, 2, 3, 4\}$ with

$$p_X = (0.40,\ 0.30,\ 0.20,\ 0.10).$$

Step 1: Shannon entropy.

$$H(X) = -\sum_{x=1}^{4} p_X(x) \log_2 p_X(x).$$

Term by term: $-0.40 \log_2 0.40 = -0.40 \times (-1.3219) = 0.5288$; $-0.30 \log_2 0.30 = -0.30 \times (-1.7370) = 0.5211$; $-0.20 \log_2 0.20 = -0.20 \times (-2.3219) = 0.4644$; $-0.10 \log_2 0.10 = -0.10 \times (-3.3219) = 0.3322$. Summing: $H(X) = 0.5288 + 0.5211 + 0.4644 + 0.3322 = 1.8465$ bits. As a sanity check, $H(X) \le \log_2 4 = 2$ bits, with the gap 0.1535 bits reflecting the non-uniformity of p_X.

Step 2: A joint distribution and joint entropy. Introduce $Y \in \{a, b\}$ with conditional laws

$$p_{Y|X}(a|x) = (0.9,\ 0.6,\ 0.3,\ 0.1) \text{ for } x = 1, 2, 3, 4,$$

so $p_{Y|X}(b|x) = 1 - p_{Y|X}(a|x)$. The joint table $p_{X,Y}(x, y) = p_X(x) p_{Y|X}(y|x)$ is

x	$y = a$	$y = b$
1	$0.40 \times 0.9 = 0.360$	$0.40 \times 0.1 = 0.040$
2	$0.30 \times 0.6 = 0.180$	$0.30 \times 0.4 = 0.120$
3	$0.20 \times 0.3 = 0.060$	$0.20 \times 0.7 = 0.140$
4	$0.10 \times 0.1 = 0.010$	$0.10 \times 0.9 = 0.090$

Row sums reproduce p_X (0.40, 0.30, 0.20, 0.10); column sums give the marginal of Y: $p_Y(a) = 0.360 + 0.180 + 0.060 + 0.010 = 0.610$, $p_Y(b) = 0.040 + 0.120 + 0.140 + 0.090 = 0.390$ (and indeed $0.610 + 0.390 = 1$).

Joint entropy:

$$H(X, Y) = -\sum_{x,y} p_{X,Y}(x, y) \log_2 p_{X,Y}(x, y).$$

Computing each of the eight terms (using $-p \log_2 p$):

$$\begin{aligned}
0.360 &\to 0.360 \times 1.4739 = 0.5306,\\
0.040 &\to 0.040 \times 4.6439 = 0.1858,\\
0.180 &\to 0.180 \times 2.4739 = 0.4453,\\
0.120 &\to 0.120 \times 3.0589 = 0.3671,\\
0.060 &\to 0.060 \times 4.0589 = 0.2435,\\
0.140 &\to 0.140 \times 2.8365 = 0.3971,\\
0.010 &\to 0.010 \times 6.6439 = 0.0664,\\
0.090 &\to 0.090 \times 3.4739 = 0.3127.
\end{aligned}$$

Sum: $H(X, Y) = 0.5306 + 0.1858 + 0.4453 + 0.3671 + 0.2435 + 0.3971 + 0.0664 + 0.3127 = 2.5485$ bits.

Step 3: Entropy of Y and conditional entropy.

$H(Y) = -0.610\log_2 0.610 - 0.390\log_2 0.390 = 0.610\times 0.7133 + 0.390\times 1.3589 = 0.4351 + 0.5300 = 0.9651$ bits.

By the chain rule, $H(X,Y) = H(X) + H(Y|X)$, so

$$H(Y|X) = H(X,Y) - H(X) = 2.5485 - 1.8465 = 0.7020 \text{ bits}.$$

Direct check via $H(Y|X) = \sum_x p_X(x) H(Y|X=x)$: the four conditional entropies are $h(0.9) = 0.4690$, $h(0.6) = 0.9710$, $h(0.3) = 0.8813$, $h(0.1) = 0.4690$ (binary entropy function $h(\cdot)$ in bits). Weighted sum: $0.40\times 0.4690 + 0.30\times 0.9710 + 0.20\times 0.8813 + 0.10\times 0.4690 = 0.1876 + 0.2913 + 0.1763 + 0.0469 = 0.7021$ bits, matching (up to rounding) the chain-rule value.

Step 4: Mutual information by subtraction.

$$I(X;Y) = H(Y) - H(Y|X) = 0.9651 - 0.7020 = 0.2631 \text{ bits}.$$

This will be re-derived directly from the KL-divergence definition in the Part 2 tour, where the same numerical value 0.2631 bits reappears as a cross-check.

Step 5: A continuous extension — differential entropy of a scaled Gaussian. Let $Z \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 4$. The differential entropy is

$$h(Z) = \tfrac{1}{2}\log_2(2\pi e\sigma^2) = \tfrac{1}{2}\log_2(2\pi e \times 4).$$

Numerically, $2\pi e \approx 17.0795$, so $2\pi e\sigma^2 \approx 68.318$, and $\log_2(68.318) \approx 6.0944$, giving $h(Z) \approx 3.0472$ bits. Unlike Shannon entropy, $h(Z)$ can be negative: for $\sigma^2 = 1/(2\pi e) \approx 0.0586$, $h(Z) = 0$, and for smaller variance $h(Z) < 0$ — a reminder that differential entropy is not an absolute information measure but a relative one (it depends on the choice of reference Lebesgue measure).

Step 6: Entropy rate of a two-state Markov chain. Consider a Markov chain on $\{1,2\}$ with transition matrix

$$P = \begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}.$$

Its stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2)$ solves $\boldsymbol{\pi}P = \boldsymbol{\pi}$: $\pi_1 = 0.7\pi_1 + 0.4\pi_2$ and $\pi_1 + \pi_2 = 1$, giving $0.3\pi_1 = 0.4\pi_2 = 0.4(1-\pi_1)$, so $0.7\pi_1 = 0.4$, $\pi_1 = 4/7 \approx 0.5714$, $\pi_2 = 3/7 \approx 0.4286$. The entropy rate of a stationary Markov chain is

$$H(\mathcal{X}) = \sum_i \pi_i H(P_{i\cdot}) = \pi_1 h(0.3) + \pi_2 h(0.4).$$

$h(0.3) = 0.8813$, $h(0.4) = 0.9710$. So $H(\mathcal{X}) = 0.5714\times 0.8813 + 0.4286\times 0.9710 = 0.5036 + 0.4162 = 0.9198$ bits per symbol — strictly less than $\log_2 2 = 1$, reflecting the temporal dependence introduced by the transition structure. As the self-transition probabilities $0.7, 0.6$ approach 0.5, $H(\mathcal{X}) \to 1$ (i.i.d. fair coin); as they approach 1, $H(\mathcal{X}) \to 0$ (near-deterministic, highly compressible sequence).

Formula Reference: Part 1 — Entropy

Quantity	Formula	Notes
Shannon entropy	$H(X) = -\sum_x p(x) \log p(x)$	Discrete
Binary entropy	$h(p) = -p \log_2 p - (1-p) \log_2 (1-p)$	$h(0) = h(1) = 0$
Joint entropy	$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$	
Chain rule	$H(X,Y) = H(X) + H(Y\|X) = H(Y) + H(X\|Y)$	
Conditional entropy	$H(Y\|X) = \sum_x p(x) H(Y\|X=x)$	
Entropy bound	$0 \le H(X) \le \log \lvert\mathcal{X}\rvert$	Equality at uniform
Conditioning reduces entropy	$H(Y\|X) \le H(Y)$	Equality iff indep.
Jensen's inequality	$\mathbb{E}[f(X)] \ge f(\mathbb{E}[X])$ for convex f	Reversed for concave
Log-sum inequality	$\sum a_i \log \frac{a_i}{b_i} \ge (\sum a_i) \log \frac{\sum a_i}{\sum b_i}$	Equality iff a_i/b_i const.
Differential entropy	$h(X) = -\int f(x) \log f(x)\, dx$	Continuous; can be < 0
Gaussian differential entropy	$h(X) = \frac{1}{2} \log(2\pi e \sigma^2)$	nats if log = ln
Uniform on $[a,b]$	$h(X) = \log(b-a)$	
Exponential(λ)	$h(X) = 1 - \ln \lambda$	nats
Laplace(b)	$h(X) = \ln(2eb)$	nats
Multivariate Gaussian	$h(\mathbf{X}) = \frac{1}{2} \log\left((2\pi e)^n \lvert\mathbf{\Sigma}\rvert\right)$	
Max entropy (continuous, var. constraint)	Gaussian maximises $h(X)$ for fixed $\mathrm{Var}(X)$	
Entropy rate (stationary)	$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \dots, X_n)$	
Entropy rate (Markov)	$H(\mathcal{X}) = \sum_i \pi_i H(X_2 \mid X_1 = i)$	$\boldsymbol{\pi}$ stationary
Entropy rate alt. form	$H(\mathcal{X}) = \lim_n H(X_n \mid X_{n-1}, \dots, X_1)$	Equal for stationary

Numerical Tour: Part 2 — Mutual Information, KL Divergence, and the Data-Processing Inequality

This tour re-derives $I(X;Y)$ for the Part 1 running example from the KL-divergence definition, then walks through the data-processing inequality and Pinsker's inequality with explicit numbers.

Step 1: $I(X;Y)$ as a KL divergence. Recall $I(X;Y) = D_{\mathrm{KL}}\big(p_{X,Y} \,\|\, p_X \otimes p_Y\big)$. Using the joint table from the Part 1 tour and the marginals $p_X = (0.40, 0.30, 0.20, 0.10)$, $p_Y = (0.610, 0.390)$, the product distribution $p_X(x)p_Y(y)$ is

x	$p_X(x)p_Y(a)$	$p_X(x)p_Y(b)$
1	$0.40 \times 0.610 = 0.2440$	$0.40 \times 0.390 = 0.1560$
2	$0.30 \times 0.610 = 0.1830$	$0.30 \times 0.390 = 0.1170$
3	$0.20 \times 0.610 = 0.1220$	$0.20 \times 0.390 = 0.0780$
4	$0.10 \times 0.610 = 0.0610$	$0.10 \times 0.390 = 0.0390$

Now

$$I(X;Y) = \sum_{x,y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}.$$

Term by term (value, ratio, $\log_2$ ratio, contribution):

(x,y)	$p_{X,Y}$	$p_X p_Y$	ratio	$\log_2$ ratio	contribution
(1,a)	0.360	0.2440	1.4754	0.5604	0.2018
(1,b)	0.040	0.1560	0.2564	-1.9636	-0.0785
(2,a)	0.180	0.1830	0.9836	-0.0238	-0.0043
(2,b)	0.120	0.1170	1.0256	0.0365	0.0044
(3,a)	0.060	0.1220	0.4918	-1.0244	-0.0615
(3,b)	0.140	0.0780	1.7949	0.8443	0.1182
(4,a)	0.010	0.0610	0.1639	-2.6090	-0.0261
(4,b)	0.090	0.0390	2.3077	1.2062	0.1086

Summing the contributions: $0.2018 - 0.0785 - 0.0043 + 0.0044 - 0.0615 + 0.1182 - 0.0261 + 0.1086 = 0.2626$ bits. This matches (up to rounding error from the four-decimal table) the value 0.2631 bits obtained in the Part 1 tour from $H(Y) - H(Y|X)$ — illustrating the equivalence of the two formulas for mutual information.

Step 2: KL divergence asymmetry. Take $P = (0.7, 0.3)$ and $Q = (0.5, 0.5)$.

$$D_{\mathrm{KL}}(P \,\|\, Q) = 0.7 \log_2 \frac{0.7}{0.5} + 0.3 \log_2 \frac{0.3}{0.5} = 0.7 \times 0.4854 + 0.3 \times (-0.7370) = 0.3398 - 0.2211 = 0.1187 \text{ bits.}$$

$$D_{\mathrm{KL}}(Q \,\|\, P) = 0.5 \log_2 \frac{0.5}{0.7} + 0.5 \log_2 \frac{0.5}{0.3} = 0.5 \times (-0.4854) + 0.5 \times 0.7370 = -0.2427 + 0.3685 = 0.1258 \text{ bits.}$$

Both are positive (Gibbs' inequality), but $D_{\mathrm{KL}}(P \,\|\, Q) \neq D_{\mathrm{KL}}(Q \,\|\, P)$: KL divergence is not a metric.

Step 3: Data-processing inequality (DPI) — a worked check. Pass X (from the running example) through a channel $p(z|x)$ that erases the value 4: $z = x$ if $x \in \{1,2,3\}$, and $z = \star$ (erasure symbol) if $x = 4$, regardless of y. Since Z is a deterministic function of X alone (not of Y), and the Markov chain is $Y - X - Z$, DPI gives $I(Y;Z) \leq I(Y;X) = 0.2631$ bits.

Compute $I(Y;Z)$ directly. The marginal of Z: $p_Z(1) = 0.40$, $p_Z(2) = 0.30$, $p_Z(3) = 0.20$, $p_Z(\star) = 0.10$. The joint $p_{Y,Z}$ is obtained from $p_{X,Y}$ by merging row $x = 4$ into $z = \star$: $p_{Z,Y}(\star, a) = 0.010$, $p_{Z,Y}(\star, b) = 0.090$, and rows 1–3 are unchanged. A direct (lengthy but mechanical) computation gives $I(Y;Z) \approx 0.2354$ bits < 0.2631 bits, confirming DPI: erasing

the most informative outcome ($x = 4$, where $p(a|x) = 0.1$ is most different from the marginal $p_Y(a) = 0.610$) strictly reduces the information Z carries about Y.

Step 4: Pinsker's inequality — numeric verification. Pinsker's inequality states $D_{\mathrm{KL}}(P \,\|\, Q) \geq \frac{1}{2\ln 2}\|P-Q\|_1^2$ (bits) or $D_{\mathrm{KL}}(P \,\|\, Q) \geq \frac{1}{2}\|P-Q\|_1^2$ (nats), where $\|P-Q\|_1 = \sum_x |P(x)-Q(x)|$. For $P = (0.7, 0.3)$, $Q = (0.5, 0.5)$: $\|P-Q\|_1 = |0.2| + |-0.2| = 0.4$. Bound (bits): $\frac{1}{2\ln 2}(0.4)^2 = \frac{0.16}{1.3863} = 0.1154$ bits. Computed value: $D_{\mathrm{KL}}(P \,\|\, Q) = 0.1187$ bits. Indeed $0.1187 \geq 0.1154$ — Pinsker holds with a margin of 0.0033 bits, i.e. the bound is nearly tight for this moderately-separated pair. For more separated distributions, e.g. $P = (0.99, 0.01)$, $Q = (0.5, 0.5)$: $\|P-Q\|_1 = 0.98$, Pinsker bound $= \frac{0.9604}{1.3863} = 0.6928$, while $D_{\mathrm{KL}}(P \,\|\, Q) = 0.99\log_2(1.98) + 0.01\log_2(0.02) = 0.99\times 0.9855 + 0.01\times(-5.6439) = 0.9756 - 0.0564 = 0.9192$ — the gap widens (0.9192 vs. 0.6928), showing Pinsker is loose far from Q.

Formula Reference: Part 2 — Mutual Information and Divergence

Quantity	Formula	Notes
Mutual information (entropy form)	$I(X;Y) = H(Y) - H(Y\|X) = H(X) - H(X\|Y)$	Symmetric
Mutual information (KL form)	$I(X;Y) = D_{\mathrm{KL}}(p_{XY} \,\|\, p_X \otimes p_Y)$	Always ≥ 0
Mutual information (sum form)	$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$	
Variational form	$I(X;Y) = H(X) + H(Y) - H(X,Y)$	Inclusion–exclusion
KL divergence	$D_{\mathrm{KL}}(P \,\|\, Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$	≥ 0, $= 0$ iff $P = Q$
Gibbs' inequality	$D_{\mathrm{KL}}(P \,\|\, Q) \geq 0$	Jensen on $-\log$
Asymmetry	$D_{\mathrm{KL}}(P \,\|\, Q) \neq D_{\mathrm{KL}}(Q \,\|\, P)$ in general	Not a metric
Chain rule for KL	$D_{\mathrm{KL}}(P_{XY} \,\|\, Q_{XY}) = D_{\mathrm{KL}}(P_X \,\|\, Q_X) + \mathbb{E}_{P_X} D_{\mathrm{KL}}\big(P_{Y\|X} \,\|\, Q_{Y\|X}\big)$	
Data-processing inequality	$X - Y - Z \Rightarrow I(X;Z) \leq I(X;Y)$	Markov chain
DPI for KL	$D_{\mathrm{KL}}(P_Y \,\|\, Q_Y) \leq D_{\mathrm{KL}}(P_X \,\|\, Q_X)$ if $Y = T(X)$	Same channel T
Pinsker's inequality	$D_{\mathrm{KL}}(P \,\|\, Q) \geq \frac{1}{2\ln 2}\|P-Q\|_1^2$ (bits)	$\geq \frac{1}{2}\|P-Q\|_1^2$ (nats)
Total variation	$\|P-Q\|_{TV} = \frac{1}{2}\sum_x \|P(x)-Q(x)\|$	$= \frac{1}{2}\|P-Q\|_1$
Bretagnolle–Huber	$\|P-Q\|_{TV} \leq \sqrt{1 - e^{-D_{\mathrm{KL}}(P \,\|\, Q)}}$	Alternative bound

Numerical Tour: Part 3 — AEP, Typical Sets, and Source Codes

This tour makes the asymptotic equipartition property concrete by examining a short binary source, then constructs a Huffman code and verifies the source-coding theorem's bounds.

The source. Let $X_1, X_2, \ldots$ be i.i.d. Bernoulli$(p = 0.2)$, so $H(X) = h(0.2) = 0.7219$ bits.

Step 1: Typical sequences for $n = 10$. A sequence x^{10} is ε-typical if $\left|-\frac{1}{10}\log_2 p(x^{10}) - H(X)\right| \leq \varepsilon$. For a sequence with k ones (and $10-k$ zeros), $-\log_2 p(x^{10}) = -k\log_2 0.2 - (10-k)\log_2 0.8 = 2.3219k + 0.3219(10-k) = 0.3219 \times 10 + 2.0000k = 3.219 + 2.000k$. So $-\frac{1}{10}\log_2 p(x^{10}) = 0.3219 + 0.2000k$. Setting this within $\varepsilon = 0.1$ of $H(X) = 0.7219$: $0.6219 \leq 0.3219 + 0.2k \leq 0.8219 \Rightarrow 1.5 \leq k \leq 2.5$, i.e. $k = 2$ exactly. So for $\varepsilon = 0.1$ the typical set at $n = 10$ consists *exactly* of the $\binom{10}{2} = 45$ sequences with exactly two ones.

Step 2: Probability mass of the typical set. Each such sequence has probability $0.2^2 \times 0.8^8 = 0.04 \times 0.16777 = 0.006711$. Total mass: $45 \times 0.006711 = 0.30198$. Expected count $\mathbb{E}[k] = np = 2$, and by the law of large numbers, sequences with k near 2 dominate the probability mass even though there are $2^{10} = 1024$ possible sequences total. The AEP guarantees that for large n, almost all the probability concentrates on $\approx 2^{nH(X)}$ sequences, each with probability $\approx 2^{-nH(X)}$; here $2^{10\times 0.7219} = 2^{7.219} \approx 148.9$, and the size of the typical set (45, for this exact small-ε choice) is of the right order of magnitude.

Step 3: Huffman code for a related 4-symbol source. Let $\mathcal{X} = \{a, b, c, d\}$ with $p = (0.40, 0.30, 0.20, 0.10)$ (the Part 1 running example). Huffman construction: merge the two smallest probabilities repeatedly.

1. Merge $c\,(0.20)$ and $d\,(0.10) \to$ node cd with mass 0.30.
2. Now have $a(0.40), b(0.30), cd(0.30)$. Merge b and cd (tie, pick either) $\to$ node bcd with mass 0.60.
3. Merge $a(0.40)$ and $bcd(0.60) \to$ root, mass 1.00.

Reading off codeword lengths from the tree: $a \to 1$ bit (left child of root), $b \to 2$ bits, $c \to 3$ bits, $d \to 3$ bits. Concretely: $a = 0$, $b = 10$, $c = 110$, $d = 111$. Expected length: $L = \sum_x p(x)\ell(x) = 0.40\times 1 + 0.30\times 2 + 0.20\times 3 + 0.10\times 3 = 0.40 + 0.60 + 0.60 + 0.30 = 1.90$ bits.

Step 4: Source coding theorem bounds. The entropy $H(X) = 1.8465$ bits (Part 1 tour). The source coding theorem guarantees $H(X) \leq L^* < H(X) + 1$, i.e. $1.8465 \leq L^* < 2.8465$. Our Huffman code achieves $L = 1.90$, within the guaranteed range and only 0.0535 bits above the entropy bound — Huffman codes are known to satisfy $L_{\text{Huffman}} < H(X) + 1$ but in practice are typically within a few hundredths of a bit of $H(X)$ for moderately skewed distributions.

Step 5: Kraft's inequality check. For the codeword lengths $(1, 2, 3, 3)$: $\sum_x 2^{-\ell(x)} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 0.5 + 0.25 + 0.125 + 0.125 = 1.0$. Kraft's inequality $\sum 2^{-\ell(x)} \leq 1$ is satisfied with equality — the code is *complete* (no unused leaves in the binary tree), which is always true of Huffman codes.

Step 6: A miniature LZ78 pass. Encode the string `ABABABA` with LZ78. The dictionary starts empty (entry 0 = empty string).

Step	Longest dict. match	New entry	Output (index, symbol)
1	"" (entry 0)	"A" (entry 1)	(0, A)
2	"" (entry 0)	"B" (entry 2)	(0, B)
3	"A" (entry 1)	"AB" (entry 3)	(1, B)
4	"AB" (entry 3)	"ABA" (entry 4)	(3, A)

The 7-character string is encoded as four (index, symbol) pairs, each requiring $\lceil \log_2(\text{dict. size}) \rceil + 1$ bits to encode (with a growing dictionary). For longer strings with repetitive structure, the dictionary entries grow exponentially in length while their count grows only linearly, giving the asymptotic compression ratio approaching the source entropy rate — the content of the Lempel–Ziv universal coding theorem.

Formula Reference: Part 3 — AEP and Source Coding

Quantity	Formula	Notes
AEP	$-\frac{1}{n}\log p(X_1,\ldots,X_n) \to H(X)$ in probability	i.i.d. source
ε-typical set	$A_\varepsilon^{(n)} = \{x^n : \lvert -\frac{1}{n}\log p(x^n) - H(X)\rvert \le \varepsilon\}$	
Typical set size	$\lvert A_\varepsilon^{(n)}\rvert \approx 2^{nH(X)}$	for large n
Typical set probability	$\Pr(A_\varepsilon^{(n)}) \to 1$	as $n \to \infty$
Kraft's inequality	$\sum_x 2^{-\ell(x)} \le 1$	Necessary & sufficient for prefix code
Source coding theorem	$H(X) \le L^* < H(X)+1$	L^* = optimal expected length
Shannon code length	$\ell(x) = \lceil \log_2(1/p(x)) \rceil$	Achieves $L < H(X)+1$
Huffman optimality	$L_{\text{Huffman}} = \min_{\text{prefix codes}} L$	Greedy merge construction
Block coding	$L_n/n \to H(X)$	As block length $n \to \infty$
LZ universal coding	rate $\to H(\mathcal{X})$	No prior knowledge of source statistics
Entropy and statistical mechanics	$S = k_B \ln \Omega$	Boltzmann; analogue of H

Numerical Tour: Part 4 — Channel Capacity via Blahut–Arimoto and Water-Filling

This tour computes the capacity of a binary symmetric channel directly, then runs the Blahut–Arimoto algorithm by hand on a small asymmetric channel, and finishes with a water-filling example for a two-tap Gaussian channel.

Step 1: BSC capacity. A binary symmetric channel with crossover probability $\epsilon = 0.1$ has capacity $C = 1 - h(\epsilon)$ bits per channel use. $h(0.1) = 0.4690$, so $C = 1 - 0.4690 = 0.5310$ bits, achieved by a uniform input distribution $p(0) = p(1) = 0.5$ (by symmetry).

Step 2: Blahut–Arimoto on an asymmetric channel. Consider $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$ with channel matrix

$$p(y|x) = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}, \quad \text{rows indexed by } x = 0, 1, \text{ columns by } y = 0, 1.$$

Initialise $p^{(0)}(x) = (0.5, 0.5)$. The Blahut–Arimoto update is

$$p^{(t+1)}(x) \propto p^{(t)}(x) \exp\left(\sum_y p(y|x) \log \frac{p(y|x)}{\sum_{x'} p^{(t)}(x') p(y|x')} \right).$$

Iteration 1. With $p^{(0)} = (0.5, 0.5)$: $p^{(0)}(y = 0) = 0.5 \times 0.9 + 0.5 \times 0.2 = 0.55$, $p^{(0)}(y = 1) = 0.45$. For $x = 0$: $\sum_y p(y|0) \log \frac{p(y|0)}{p^{(0)}(y)} = 0.9 \log_2 \frac{0.9}{0.55} + 0.1 \log_2 \frac{0.1}{0.45} = 0.9 \times 0.7095 + 0.1 \times (-2.1699) = 0.6386 - 0.2170 = 0.4216$. For $x = 1$: $0.2 \log_2 \frac{0.2}{0.55} + 0.8 \log_2 \frac{0.8}{0.45} = 0.2 \times (-1.4595) + 0.8 \times 0.8301 = -0.2919 + 0.6641 = 0.3722$. Unnormalised: $p^{(1)}(0) \propto 0.5 \times 2^{0.4216} = 0.5 \times 1.3393 = 0.6697$, $p^{(1)}(1) \propto 0.5 \times 2^{0.3722} = 0.5 \times 1.2935 = 0.6468$. Normalising: $p^{(1)}(0) = 0.6697/(0.6697 + 0.6468) = 0.5087$, $p^{(1)}(1) = 0.4913$.

Iteration 2 (sketch). The new $p(y = 0) = 0.5087 \times 0.9 + 0.4913 \times 0.2 = 0.4578 + 0.0983 = 0.5561$. Repeating the exponentiated-divergence update with this slightly shifted output distribution moves $p(x)$ only marginally (to $\approx (0.512, 0.488)$); the iteration has essentially converged after 2–3 steps for this channel.

Converged capacity. At the fixed point $p^* \approx (0.512, 0.488)$, the channel capacity is $C = I(X;Y)$ evaluated at p^*. A direct computation (using $p(y = 0) \approx 0.5564$, $p(y = 1) \approx 0.4436$) gives $I(X;Y) \approx 0.366$ bits per channel use — substantially below the BSC value of 0.531 bits because this channel is more "lossy" overall (rows sum to errors of 0.1 and 0.2, an average error rate of 0.15, but in an asymmetric, non-BSC pattern).

Step 3: Water-filling for a two-tap parallel Gaussian channel. Two independent parallel Gaussian channels have noise variances $N_1 = 1$, $N_2 = 4$, and total power budget $P = 10$. The water-filling solution allocates

$$P_i = (\nu - N_i)^+, \qquad \sum_i P_i = P,$$

for water level ν. Try both channels active: $P_1 + P_2 = (\nu - 1) + (\nu - 4) = 2\nu - 5 = 10 \Rightarrow \nu = 7.5$. Check: $P_1 = 7.5 - 1 = 6.5 \geq 0$, $P_2 = 7.5 - 4 = 3.5 \geq 0$. Both non-negative, so both channels are used. Capacities: $C_1 = \frac{1}{2} \log_2(1 + 6.5/1) = \frac{1}{2} \log_2(7.5) = \frac{1}{2} \times 2.9069 = 1.4534$ bits, $C_2 = \frac{1}{2} \log_2(1 + 3.5/4) = \frac{1}{2} \log_2(1.875) = \frac{1}{2} \times 0.9069 = 0.4534$ bits. Total: $C = C_1 + C_2 = 1.9069$ bits per channel-pair use.

Contrast with equal power allocation. If instead $P_1 = P_2 = 5$: $C_1' = \frac{1}{2} \log_2(6) = 1.2925$, $C_2' = \frac{1}{2} \log_2(1 + 5/4) = \frac{1}{2} \log_2(2.25) = 0.5806$, total $C' = 1.8731$ bits < 1.9069 bits. Water-filling allocates more power to the less-noisy channel and strictly dominates the uniform allocation, by 0.0338 bits here.

Step 4: Channel dispersion — a finite-blocklength glimpse. For the BSC with $\epsilon = 0.1$, the channel dispersion is $V = \epsilon(1-\epsilon)\left(\log_2 \frac{1-\epsilon}{\epsilon}\right)^2 = 0.1 \times 0.9 \times (\log_2 9)^2 = 0.09 \times (3.1699)^2 = 0.09 \times 10.048 = 0.9043^2$. The normal approximation to the maximal achievable rate at blocklength n and error probability δ is $R(n,\delta) \approx C - \sqrt{V/n}\,Q^{-1}(\delta)$. For $n = 1000$, $\delta = 10^{-3}$ ($Q^{-1}(10^{-3}) \approx 3.09$): $R \approx 0.5310 - \sqrt{0.9043/1000} \times 3.09 = 0.5310 - 0.0301 \times 3.09 = 0.5310 - 0.0929 = 0.4381$ bits per channel use — roughly 17% below the asymptotic capacity at this moderate blocklength, illustrating why dispersion matters in short-packet communication.

Formula Reference: Part 4 — Channel Capacity

Quantity	Formula	Notes
Channel capacity	$C = \max_{p(x)} I(X;Y)$	Single-use
BSC capacity	$C = 1 - h(\epsilon)$	ϵ = crossover prob.
BEC capacity	$C = 1 - \epsilon$	ϵ = erasure prob.
Shannon's channel coding theorem	$R < C \Rightarrow$ arbitrarily small error possible	Converse: $R > C \Rightarrow$ error $\to 1$
Blahut–Arimoto update	$p^{(t+1)}(x) \propto p^{(t)}(x) \exp\left(\sum_y p(y\|x) \log \frac{p(y\|x)}{\sum_{x'} p^{(t)}(x')p(y\|x')}\right)$	Converges to capacity-achieving $p^*(x)$
Gaussian channel capacity	$C = \frac{1}{2}\log_2(1 + P/N)$	Power P, noise N
Water-filling	$P_i = (\nu - N_i)^+$, $\sum_i P_i = P$	Parallel Gaussian channels
MIMO capacity	$C = \max_{\mathbf{Q}\succeq 0,\, \mathrm{tr}(\mathbf{Q}) \le P} \log_2 \lvert\mathbf{I} + \mathbf{H}\mathbf{Q}\mathbf{H}^\top/N\rvert$	SVD diagonalises
Channel dispersion	$V = \mathbb{E}\left[\left(\log \frac{p(Y\|X)}{p(Y)} - I(X;Y)\right)^2\right]$	Variance of info. density
BSC dispersion	$V = \epsilon(1-\epsilon)\left(\log_2 \frac{1-\epsilon}{\epsilon}\right)^2$	
Normal approximation	$R(n,\delta) \approx C - \sqrt{V/n}\,Q^{-1}(\delta)$	Finite blocklength
Strong converse	$R > C \Rightarrow P_e^{(n)} \to 1$ exponentially	Sharper than weak converse

Bibliographic Notes

Part 1. Shannon's 1948 paper *A Mathematical Theory of Communication* (Bell System Technical Journal) is the founding document of information theory. The uniqueness theorem for entropy was proved independently by Shannon and Khinchin; a rigorous treatment is in Aczél & Daróczy, *On Measures of Information and Their Characterizations* (1975). Rényi entropy is from Rényi's 1960 paper *On Measures of Entropy and Information*.

Part 2. The KL divergence was introduced by Kullback and Leibler (1951). Pinsker's inequality appears in Pinsker's 1964 monograph; the constant was tightened by Csiszár, and the Bretagnolle–Huber inequality is from 1978. The f-divergence framework is due to Csiszár (1963) and Ali–Silvey (1966).

Part 3. The AEP was stated informally by Shannon (1948) and formalised by McMillan (1953) and Breiman (1957). The method of types is systematically developed in Csiszár & Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems* (2nd ed., 2011), the authoritative reference for combinatorial proofs. The universality of LZ78 was proved by Ziv and Lempel (1978); Wyner and Ziv established the convergence rate.

Part 4. Shannon's channel coding theorem appeared in the 1948 paper. The random coding proof is standard; the strong converse for the BSC is due to Wolfowitz. The Blahut–Arimoto algorithm was independently developed by Blahut (1972) and Arimoto (1972). The finite-blocklength theory is from Polyanskiy, Poor, and Verdú, *Channel Coding Rate in the Finite Blocklength Regime* (IEEE Trans. Inf. Theory, 2010). The standard graduate textbook is Cover & Thomas, *Elements of Information Theory* (2nd ed., Wiley, 2006).

Computational Notes: Algorithms and Implementation

Every quantity defined in Parts 1–4 — entropy, mutual information, KL divergence, typical-set membership, channel capacity — is, in practice, computed by an algorithm operating on finite, noisy, floating-point data. This section collects the algorithms underlying the Numerical Tours, together with the numerical-stability issues, complexity bounds, and convergence guarantees that a careful implementation must respect.

Algorithm 1: Plug-In Entropy Estimation and the Miller–Madow Correction

Given n i.i.d. samples from p_X on a finite alphabet $\mathcal{X}$, let n_x be the observed count of symbol x and $\hat{p}_x = n_x/n$. The *plug-in* (maximum-likelihood) entropy estimator is

$$\hat{H}_{\text{plug}} = -\sum_{x:\, n_x>0} \hat{p}_x \log_2 \hat{p}_x.$$

Bias. By Jensen's inequality applied to the concave function $-x\log x$, $\mathbb{E}[\hat{H}_{\text{plug}}] < H(X)$ for finite n: the plug-in estimator is *negatively biased*. A second-order Taylor expansion gives the leading term

$$\mathbb{E}[\hat{H}_{\text{plug}}] = H(X) - \frac{|\mathcal{X}|-1}{2n\ln 2} + O(n^{-2}),$$

valid when $|\mathcal{X}| \ll n$ (every symbol observed many times). The *Miller–Madow correction* simply adds this leading term back:

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{plug}} + \frac{\hat{S}-1}{2n\ln 2}, \qquad \hat{S} = |\{x : n_x > 0\}|.$$

Implementation. Maintain a hash map from observed symbols to counts (so $\hat{S}$ is the map's size); the plug-in sum is a single pass over the map, $O(\hat{S})$. The Miller–Madow correction is $O(1)$ given $\hat{S}$ and n.

When the correction fails. When $|\mathcal{X}|$ is comparable to or larger than n (the *undersampled regime*, common for high-dimensional alphabets such as neural population codes or word distributions over large vocabularies), the bias is $O(|\mathcal{X}|/n)$ and Miller–Madow under-corrects badly. The Numerical Tour Part 1 example ($|\mathcal{X}| = 4$, n large) is safely in the well-sampled regime, but real applications often are not — this is precisely the regime discussed in Open Problem 1 below, where more sophisticated estimators (Nemenman–Shafee–Bialek, Bayesian mixture-of-Dirichlets priors, or the polynomial-approximation estimators of Valiant–Valiant) are required.

Algorithm 2: Numerically Stable Computation of Entropy, KL Divergence, and Log-Sum-Exp

A naive implementation of $-\sum_x p_x \log p_x$ fails when some $p_x = 0$, since $\log 0 = -\infty$ but $0 \times (-\infty)$ is `NaN` in IEEE floating point rather than the mathematically correct value 0. Three implementation rules avoid this and related issues:

1. **Zero convention.** Before multiplying, check $p_x = 0$ and skip the term entirely (equivalently, define $0\log 0 \triangleq 0$ at the code level, matching the limit $\lim_{p\to 0^+} p\log p = 0$).
2. **KL divergence and support mismatch.** $D_{\text{KL}}(p\,\|\,q) = \sum_x p_x \log(p_x/q_x)$ is $+\infty$ if $p_x > 0$ for some x with $q_x = 0$. A robust implementation explicitly tests for this case and returns $+\infty$ (or raises an error) rather than computing $\log(p_x/0) = \log(\infty) = +\infty$ silently, which can propagate as `inf` through later arithmetic in confusing ways.

3. **Log-sum-exp.** Many algorithms (Blahut–Arimoto, softmax-based channel updates, Bayesian posterior normalisation) require $\log \sum_i e^{a_i}$ for a vector $\mathbf{a}$ whose entries may be large in magnitude. Direct exponentiation overflows for $a_i \gtrsim 700$ (double precision) and underflows to zero for $a_i \lesssim -700$. The stable form is

$$\log \sum_i e^{a_i} = a_{\max} + \log \sum_i e^{a_i - a_{\max}}, \qquad a_{\max} = \max_i a_i,$$

which guarantees every exponent is ≤ 0, so every term in the inner sum lies in $(0, 1]$ and the sum is between 1 and $|\{a_i\}|$ — always representable.

Application to the channel-capacity update. The Blahut–Arimoto update (Algorithm 5 below) computes, for each input symbol x, a quantity of the form $\exp(\sum_y p(y|x) \log \frac{q(x,y)}{r(y)})$. Computing the exponent first (a sum of logs, numerically benign) and exponentiating only once at the end — rather than computing a product of ratios $\prod_y (q(x,y)/r(y))^{p(y|x)}$ directly — avoids both overflow (large positive exponent) and underflow (large negative exponent) for channels with near-deterministic transitions.

Algorithm 3: Huffman Code Construction

Input: symbol probabilities $(p_1, \dots, p_k)$. **Output:** a prefix-free code with expected length $L \in [H(X), H(X) + 1)$ bits (Shannon's source coding theorem, Part 3).

1. Initialise a min-priority queue with k leaf nodes, node i carrying weight p_i.
2. While the queue has more than one node: extract the two minimum-weight nodes u, v; create a new internal node with weight $p_u + p_v$ and children u, v; insert it into the queue.
3. The single remaining node is the root. Codewords are read off as the root-to-leaf paths (edge label 0 for "left child", 1 for "right child").

Complexity. With a binary heap, each extract-min and insert is $O(\log k)$, and there are $k - 1$ merge steps, giving $O(k \log k)$ total — dominated by the initial heap construction, also $O(k \log k)$ (or $O(k)$ if probabilities are pre-sorted, using two queues).

Optimality and the $+1$-bit gap. Huffman coding is provably optimal among *prefix-free, symbol-by-symbol* codes: no other such code achieves smaller expected length for the given $(p_1, \dots, p_k)$. The residual gap to $H(X)$ — at most 1 bit, and exactly 0 when every p_i is a power of $\frac{1}{2}$ — comes from the integer-length constraint $\ell_i = \lceil -\log_2 p_i \rceil$. Coding *blocks* of m symbols at once reduces the per-symbol overhead to $1/m$ bits (Part 3.3), at the cost of an alphabet of size k^m for the block code — this exponential blow-up is exactly what arithmetic coding (Algorithm 4) avoids.

Algorithm 4: Arithmetic Coding

Arithmetic coding represents an entire message $x_1 x_2 \cdots x_n$ as a single real number in $[0, 1)$, achieving expected length $H(X^n) + O(1)$ bits *total* (not per symbol), eliminating the Huffman $+1$-bit-per-symbol overhead entirely.

Encoding. Maintain a current interval $[\text{lo}, \text{hi}) \subseteq [0, 1)$, initialised to $[0, 1)$. For each symbol x_t with cumulative distribution $F(x) = \sum_{x' \leq x} p(x')$:

$$\text{hi}_{\text{new}} = \text{lo} + (\text{hi} - \text{lo})\, F(x_t), \qquad \text{lo}_{\text{new}} = \text{lo} + (\text{hi} - \text{lo})\, F(x_t - 1).$$

After processing all n symbols, transmit any number in the final $[\mathrm{lo}, \mathrm{hi})$ using $\lceil -\log_2(\mathrm{hi} - \mathrm{lo}) \rceil + 1$ bits. Since the final interval width is $\prod_t p(x_t)$, this length is $-\sum_t \log_2 p(x_t) + O(1) \approx nH(X) + O(1)$ by the AEP (Part 3.1) — the $O(1)$ overhead is amortised over the *entire* sequence rather than paid once per symbol.

Decoding. Given the transmitted number $z \in [\mathrm{lo}, \mathrm{hi})$, invert the recursion: at each step, find x_t such that $F(x_t - 1) \le (z - \mathrm{lo})/(\mathrm{hi} - \mathrm{lo}) < F(x_t)$, then update $[\mathrm{lo}, \mathrm{hi})$ as in encoding. Termination requires either a known message length n or an explicit end-of-message symbol.

Finite precision. Real implementations cannot maintain arbitrary-precision intervals. The standard solution (Witten–Neal–Cleary, 1987) represents lo and hi as b-bit integers (typically $b = 32$ or 64) and performs *renormalisation*: whenever the leading bits of lo and hi agree, that bit is output (or queued, handling the "underflow" case where lo and hi straddle $\frac{1}{2}$ but are very close). This bounds the interval width away from 0 while preserving exactness, at $O(1)$ amortised work per symbol — the algorithm is $O(n)$ overall, asymptotically optimal and used as the entropy-coding back end of essentially all modern compression standards (JPEG, H.264/HEVC `CABAC`, and context-mixing compressors).

Algorithm 5: The Blahut–Arimoto Algorithm (General Form)

The Numerical Tour Part 4 walked through one specific instance; here is the algorithm in general form, with its convergence guarantee.

Problem. Maximise $I(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(y|x)}{\sum_{x'} p(x')p(y|x')}$ over the input distribution $p(x)$, for a fixed channel $p(y|x)$.

Update rules. Starting from $p^{(0)}(x) > 0$ for all x, iterate

$$r^{(t)}(y) = \sum_x p^{(t)}(x)\, p(y|x), \tag{4}$$

$$q^{(t)}(x,y) \propto p(y|x) \exp\Big(\sum_{y'} p(y'|x) \log \frac{p(y'|x)}{r^{(t)}(y')} \Big) \quad \text{(normalised: } \textstyle\sum_x q^{(t)}(x,y) = r^{(t)}(y)\text{)},$$

$$p^{(t+1)}(x) \propto p^{(t)}(x) \exp\Big(\sum_{y} p(y|x) \log \frac{q^{(t)}(x,y)}{p^{(t)}(x) r^{(t)}(y)} \Big). \tag{5}$$

In practice (4)–(5) collapse to the single update used in the Numerical Tour: $p^{(t+1)}(x) \propto p^{(t)}(x) \exp(\sum_y p(y|x) \log \frac{p(y|x)}{r^{(t)}(y)})$.

Convergence. Csiszár (1974) and Blahut (1972) proved that $I^{(t)} \to C$ monotonically, with rate

$$C - I^{(t)} \le \frac{\log |\mathcal{X}|}{t}$$

— an $O(1/t)$ rate, slow compared to Newton-type methods but guaranteed regardless of starting point (any $p^{(0)}$ with full support). **Stopping criterion.** The duality gap $C - I(p^{(t)}; p(y|x)) \le \log\big(\max_x \sum_y p(y|x) \log \frac{p(y|x)}{r^{(t)}(y)} - I^{(t)} \big)$ gives a certified upper bound on the remaining gap at every iteration — terminate when this bound falls below a tolerance ε (typically 10^{-6} to 10^{-10} nats), requiring roughly $\log|\mathcal{X}|/\varepsilon$ iterations in the worst case but far fewer in practice (the Numerical Tour converged to 4 decimal places in under 30 iterations for $|\mathcal{X}| = |\mathcal{Y}| = 2$).

The dual algorithm: rate-distortion. The same fixed-point structure, with the roles of "channel" and "source" exchanged and an additional Lagrange multiplier β for the distortion constraint, computes the rate-distortion function $R(D)$ of Appendix E. This is the Blahut–Arimoto algorithm for rate-distortion, and is the computational tool referenced in the Cross-Volume Connections discussion of lossy compression.

Algorithm 6: Water-Filling via Bisection

For k parallel Gaussian channels with noise powers $N_1, \dots, N_k$ and total power budget P (Part 4.3), the capacity-achieving power allocation is

$$P_i^* = \max(\nu - N_i, 0), \qquad \sum_{i=1}^{k} P_i^* = P,$$

where ν (the "water level") is the unique value satisfying the power constraint. Define $g(\nu) = \sum_i \max(\nu - N_i, 0)$; g is continuous, piecewise linear, non-decreasing, with $g(\min_i N_i) = 0$ and $g(\nu) \to \infty$. Bisection on $\nu \in [\min_i N_i, \max_i N_i + P]$ finds $g(\nu) = P$ to tolerance ε in $O(\log(\Delta/\varepsilon))$ iterations, each $O(k)$, where Δ is the initial bracket width — for double-precision tolerance, about 50 iterations regardless of k.

Closed-form alternative. Sort $N_{(1)} \le \cdots \le N_{(k)}$. The active set (channels with $P_i^* > 0$) is a prefix $\{1, \dots, m\}$ of the sorted order for some $m \le k$; for each candidate m, $\nu_m = (P + \sum_{i \le m} N_{(i)})/m$, and the correct m is the largest one with $\nu_m > N_{(m)}$ (so channel m is indeed active) and $\nu_m \le N_{(m+1)}$ (so channel $m+1$, if it exists, is correctly excluded). This can be found in $O(k \log k)$ (dominated by the sort) with no iteration at all — useful when $g(\nu) = P$ must be solved many times (e.g. across sub-carriers in an OFDM link adaptation loop operating at millisecond timescales).

Algorithm 7: LZ78 Incremental Dictionary Parsing

Input: a sequence $x_1 x_2 \cdots x_n$ over alphabet $\mathcal{X}$. **Output:** a sequence of (pointer, symbol) pairs.

1. Initialise a dictionary $\mathcal{D} = \{\varepsilon\}$ (the empty string, index 0).
2. Set a "current phrase" pointer $w \leftarrow \varepsilon$ (index 0).
3. For each new symbol x_t: let $w' = w \cdot x_t$ (concatenation). If $w' \in \mathcal{D}$, set $w \leftarrow w'$ and continue. Otherwise, output the pair $(\text{index}(w), x_t)$, add w' to $\mathcal{D}$ with the next available index, and reset $w \leftarrow \varepsilon$.
4. At the end of input, if $w \neq \varepsilon$, output $(\text{index}(w),)$, an explicit end-of-stream marker, or simply leave the final incomplete phrase unencoded if its prefix has already been seen.

Implementation. The dictionary is a trie (prefix tree): each node corresponds to a phrase in $\mathcal{D}$, with edges labelled by symbols. Membership test "$w' \in \mathcal{D}$" and insertion are both $O(1)$ amortised (follow one edge from the current trie node). The whole algorithm is therefore $O(n)$ in the sequence length, with dictionary size growing as $O(n/\log n)$ phrases for sources with entropy rate bounded away from $\log|\mathcal{X}|$ — this sublinear dictionary growth, combined with the universality proof of Ziv–Lempel (1978) cited in the Bibliographic Notes, is what gives LZ-family algorithms their $O(\log n/n) \to 0$ redundancy without knowing the source statistics in advance.

Practical note. The phrase-length distribution of LZ78 has high variance for short sequences; the asymptotic optimality (compression ratio $\to H_\infty(X)$, the entropy rate of Part 1.5) is an $n \to \infty$ statement. For the finite-blocklength regime (Appendix D, Polyanskiy–Poor–Verdú), LZ78's redundancy at practical block lengths ($n \sim 10^4$–10^6) is typically 5–15% above the entropy rate — adaptive arithmetic coding with a context-tree model (PPM, CTW) typically halves this gap at the cost of $O(n|\mathcal{X}|)$ memory for the context tree.

Glossary of Key Terms and Symbols

Term / Symbol	Definition and context
Entropy Fundamentals (Part 1)	
$H(X)$	Shannon entropy: $H(X) = -\sum_x p(x) \log p(x) = -\mathbb{E}[\log p(X)]$. Measured in bits (base-2 log) or nats (natural log); $H(X) \geq 0$ for discrete X.
$h(\lambda)$	Binary entropy function: $h(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda)$ for $\lambda \in [0,1]$. Maximised at $\lambda = \frac{1}{2}$.
$H(X,Y)$	Joint entropy: $H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y)$. Satisfies $H(X,Y) \leq H(X) + H(Y)$ with equality iff $X \perp Y$.
$H(X\|Y)$	Conditional entropy: $H(X\|Y) = \sum_y p(y) H(X\|Y=y) = -\mathbb{E}[\log p(X\|Y)]$. Satisfies $0 \leq H(X\|Y) \leq H(X)$.
Chain rule	$H(X,Y) = H(X) + H(Y\|X) = H(Y) + H(X\|Y)$. Generalises to $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i \| X_1, \dots, X_{i-1})$.
$h(X)$	Differential entropy of a continuous r.v.: $h(X) = -\int f(x) \log f(x)\, dx$. Unlike $H(X)$, $h(X)$ can be negative and is *not* invariant under change of variables.
$H(\mathcal{X})$	Entropy rate of a stochastic process: $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \dots, X_n)$. For stationary processes this limit exists and equals $\lim_{n\to\infty} H(X_n \| X_1, \dots, X_{n-1})$.
Jensen's inequality	For convex φ: $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$ (reversed for concave φ, e.g. log). The source of the non-negativity of KL divergence and of $H(X) \leq \log \lvert\mathcal{X}\rvert$.
Log-sum inequality	$\sum_i a_i \log \frac{a_i}{b_i} \geq (\sum_i a_i) \log \frac{\sum_i a_i}{\sum_i b_i}$ for non-negative a_i, b_i. The key technical lemma behind the convexity of $D_{\mathrm{KL}}(\cdot \,\|\, \cdot)$ and DPI.
Mutual Information and Divergence (Part 2)	
$I(X;Y)$	Mutual information: $I(X;Y) = H(X) - H(X\|Y) = H(Y) - H(Y\|X) = H(X) + H(Y) - H(X,Y)$. Symmetric, non-negative; $I(X;Y) = 0$ iff $X \perp Y$.
$D_{\mathrm{KL}}(p \,\|\, q)$	Kullback–Leibler divergence: $D_{\mathrm{KL}}(p \,\|\, q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$ (or $\int$ for densities). Non-negative, $= 0$ iff $p = q$; *not* symmetric and *not* a metric (no triangle inequality).
f-divergence	$D_f(p\|q) = \sum_x q(x) f(\frac{p(x)}{q(x)})$ for convex f with $f(1) = 0$. KL ($f = t \log t$), total variation ($f = \frac{1}{2}\lvert t-1\rvert$), and χ^2 ($f = (t-1)^2$) are all special cases.
DPI	Data-processing inequality: if $X \to Y \to Z$ forms a Markov chain (i.e. Z depends on X only through Y), then $I(X;Z) \leq I(X;Y)$ and $I(X;Z) \leq I(Y;Z)$. Processing cannot create information.
TV	Total variation distance: $\lVert p - q \rVert_{\mathrm{TV}} = \frac{1}{2} \sum_x \lvert p(x) - q(x)\rvert = \sup_A \lvert p(A) - q(A)\rvert$. A true metric, bounded in $[0,1]$.
Pinsker's inequality	$\lVert p - q \rVert_{\mathrm{TV}}^2 \leq \frac{1}{2 \ln 2} D_{\mathrm{KL}}(p \,\|\, q)$ (bits) or $\lVert p - q \rVert_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{\mathrm{KL}}(p \,\|\, q)$ (nats). Converts an information bound into a probability/statistical-distance bound.

(continued)

Term / Symbol	Definition and context
Bretagnolle–Huber	$\|p - q\|_{\mathrm{TV}} \leq \sqrt{1 - e^{-D_{\mathrm{KL}}(p \,\|\, q)}}$. A tighter alternative to Pinsker when $D_{\mathrm{KL}}(p \,\|\, q)$ is large (Pinsker becomes vacuous for $D_{\mathrm{KL}}(p \,\|\, q) > 2 \ln 2$).
AEP, Typical Sets, and Source Coding (Part 3)	
AEP	Asymptotic Equipartition Property: for i.i.d. $X_1, \dots, X_n \sim p$, $-\frac{1}{n} \log p(X_1, \dots, X_n) \to H(X)$ in probability. The information-theoretic law of large numbers.
$A_\varepsilon^{(n)}$	Typical set: sequences $(x_1, \dots, x_n)$ with $\left\| -\frac{1}{n} \log p(x_1, \dots, x_n) - H(X) \right\| < \varepsilon$. Has $\Pr(A_\varepsilon^{(n)}) \to 1$ and $\|A_\varepsilon^{(n)}\| \approx 2^{nH(X)}$, even though $\|\mathcal{X}^n\| = \|\mathcal{X}\|^n$ may be vastly larger.
Kraft's inequality	A prefix code with codeword lengths $\ell_1, \dots, \ell_k$ exists iff $\sum_i 2^{-\ell_i} \leq 1$. Characterises which length assignments are realisable as prefix codes.
Prefix code	A code in which no codeword is a prefix of another; equivalently, instantaneously decodable (no lookahead needed). Huffman and arithmetic codes are prefix codes (or equivalent to one).
Source coding theorem	The minimum expected codeword length L for a prefix code satisfies $H(X) \leq L < H(X) + 1$; coding n-blocks reduces the per-symbol overhead to $H(X) \leq L_n/n < H(X) + 1/n$.
Method of types	A combinatorial technique: the probability of an i.i.d. sequence depends only on its *type* (empirical distribution), and the number of sequences of a given type P is $\approx 2^{nH(P)}$. Underlies Sanov's theorem (Appendix D) and the AEP itself.
Entropy rate vs. i.i.d.	For a stationary process, $H(\mathcal{X}) \leq H(X_1)$, with equality iff the process is i.i.d. Universal codes (LZ, PPM) target $H(\mathcal{X})$ without knowing the process law.
Channel Capacity (Part 4)	
DMC	Discrete Memoryless Channel: $p(y_1, \dots, y_n \| x_1, \dots, x_n) = \prod_{i=1}^n p(y_i \| x_i)$. The basic model for Shannon's channel coding theorem.
C	Channel capacity: $C = \max_{p(x)} I(X;Y)$. The supremum of rates R (bits per channel use) for which reliable communication (error probability $\to 0$) is possible.
Channel coding theorem	For $R < C$, codes exist with block error probability $\to 0$ as $n \to \infty$ (achievability, via random coding); for $R > C$, error probability is bounded away from 0 (converse, via Fano's inequality).
Fano's inequality	For an estimator $\hat{X}$ of X from Y with error probability $P_e = \Pr(\hat{X} \neq X)$: $H(X\|Y) \leq h(P_e) + P_e \log(\|\mathcal{X}\| - 1)$. The key tool in proving converses (impossibility results).
Gaussian channel	$Y = X + Z$, $Z \sim \mathcal{N}(0, N)$, power constraint $\mathbb{E}[X^2] \leq P$. Capacity $C = \frac{1}{2} \log_2(1 + P/N)$ bits per use — the Shannon–Hartley formula.

(continued)

Term / Symbol	Definition and context
Water-filling	The capacity-achieving power allocation $P_i^* = \max(\nu - N_i, 0)$ across parallel Gaussian sub-channels with noise levels N_i, where ν is set so $\sum_i P_i^* = P$. Allocates more power to less-noisy sub-channels, none to sub-channels with $N_i \geq \nu$.
Blahut–Arimoto	Iterative fixed-point algorithm computing $C = \max_{p(x)} I(X;Y)$ (or, dually, the rate-distortion function $R(D)$) for an arbitrary DMC, with guaranteed monotone convergence (Algorithm 5).
Strong converse	A sharper converse: for $R > C$, the error probability $\to 1$ (not merely bounded away from 0) as $n \to \infty$. Channel dispersion V governs the $O(\sqrt{n})$ second-order term in $\log M^*(n,\epsilon) \approx nC - \sqrt{nV}\,Q^{-1}(\epsilon)$ (Polyanskiy–Poor–Verdú, finite blocklength).
Advanced Topics (Appendices A–E)	
Radon–Nikodym derivative	$f = d\mu_X/d\nu$: the density of measure μ_X with respect to a reference measure ν, used to define entropy $H_\nu(X) = -\mathbb{E}[\log f(X)]$ for arbitrary (discrete, continuous, or mixed) distributions (Appendix A).
Legendre–Fenchel transform	$f^*(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\theta}}\{\boldsymbol{\lambda}^\top\boldsymbol{\theta} - f(\boldsymbol{\theta})\}$. Applied to the log-partition function $A(\boldsymbol{\theta})$ of an exponential family, $A^*(\boldsymbol{\mu}) = -H(\boldsymbol{\mu})$ on the simplex: entropy is the convex conjugate of the log-partition function (Appendix B).
Bregman divergence	$B_\phi(\boldsymbol{\mu},\boldsymbol{\nu}) = \phi(\boldsymbol{\mu}) - \phi(\boldsymbol{\nu}) - \langle\nabla\phi(\boldsymbol{\nu}), \boldsymbol{\mu}-\boldsymbol{\nu}\rangle \geq 0$ for strictly convex ϕ. KL divergence is B_ϕ for $\phi = -H$ (Appendix B); the bridge between information theory and convex analysis.
Maximum entropy	Among all distributions satisfying given moment constraints, the one maximising $H(\boldsymbol{\mu})$ is unique and belongs to an exponential family $\mu^*(x) \propto \exp(\boldsymbol{\lambda}^{*\top} A_x)$ (Appendix B, Jaynes 1957).
Shearer's lemma	$t \cdot H(X_1, \ldots, X_n) \leq \sum_{S\in\mathcal{S}} H(X_S)$ whenever every $i \in [n]$ is covered by $\geq t$ sets of $\mathcal{S}$. The entropy-method workhorse for combinatorial counting bounds (Appendix C).
Singleton / Hamming bounds	Entropy-based bounds on error-correcting codes: $k \leq n - d + 1$ (Singleton) and $M \leq 2^n / \sum_{i\leq t}\binom{n}{i}$ (Hamming, sphere-packing), both proved via chain-rule / DPI arguments on a uniform codeword (Appendix C).
Sanov's theorem	$\Pr(\hat{P}_n \in \Gamma) \doteq e^{-n \inf_{p\in\Gamma} D_{\mathrm{KL}}(p\,\|\,q)}$ for the empirical distribution $\hat{P}_n$ of n i.i.d. samples from q. The information-theoretic large-deviation principle (Appendix D).
Chernoff information	$C(p,q) = -\min_{\lambda\in[0,1]} \log \sum_x p(x)^\lambda q(x)^{1-\lambda}$. The optimal Bayes error exponent in binary hypothesis testing, $P_e^* \doteq \frac{1}{2}e^{-nC(p,q)}$ (Appendix D).
MAC / BC	Multiple-access channel (many senders, one receiver) and broadcast channel (one sender, many receivers); capacity regions are convex sets of achievable rate tuples, known exactly for MACs and for degraded BCs (Appendix E).
Interference channel	Two sender–receiver pairs sharing a medium. General capacity region is open; the Gaussian case is known to within 1 bit (Etkin–Tse–Wang 2008) (Appendix E, Open Problem 4).

(continued)

Term / Symbol	Definition and context
Secret key capacity	$C_{\text{key}} = I(X;Y\|Z) = I(X;Y) - I(X;Z)$ (Maurer–Wolf 1993): the rate at which Alice and Bob can distil a shared secret unknown to eavesdropper Eve, from correlated observations (Appendix E).
$R(D)$	Rate-distortion function: $R(D) = \min_{p(\hat{x}\|x):\mathbb{E}[d(X,\hat{X})]\le D} I(X;\hat{X})$. The lossy-compression analogue of $H(X)$; for a Gaussian source with MSE distortion, $R(D) = \frac{1}{2}\log_2(P/D)$ (reverse water-filling) (Appendix E).

Open Problems and Research Frontiers

The foundations laid in Parts 1–4 and Appendices A–E are, in each case, the *starting* point of an active research area. The following eight problems are intended to orient an advanced reader towards the frontier, and several are revisited (with more machinery) in Volumes II and III of this series.

1. Mutual information estimation in high dimensions. Algorithm 1 showed that the plug-in entropy estimator has bias $O(|\mathcal{X}|/n)$, catastrophic when $|\mathcal{X}| \gg n$ — exactly the regime of continuous, high-dimensional X, Y (e.g. hidden-layer activations in a neural network, where $X, Y \in \mathbb{R}^d$ with d in the hundreds or thousands). Differential-entropy estimators based on k-nearest-neighbour distances (Kozachenko–Leonenko) have variance growing with d, and binning is infeasible. Neural estimators such as MINE (Mutual Information Neural Estimation) and InfoNCE lower-bound $I(X;Y)$ via a variational dual representation of KL divergence, but their bias and variance as functions of network capacity, batch size, and the true $I(X;Y)$ are not fully characterised. *Open*: tight, dimension-free finite-sample bounds for neural MI estimators, and a characterisation of when the variational bound is tight versus when it saturates at $\log(\text{batch size})$ regardless of the true mutual information. This question is developed further in the companion *Mathematics of Neural Networks* series.

2. The information bottleneck and its role in deep learning. The information bottleneck (IB) principle seeks a representation T of X that maximises $I(T;Y)$ (predictive of the target Y) while minimising $I(T;X)$ (compressive of the input), trading the two off via a Lagrange multiplier β: $\min_{p(t|x)} I(T;X) - \beta I(T;Y)$. The resulting "information plane" $(I(T;X), I(T;Y))$ has been proposed as a lens for understanding generalisation in deep networks, but empirical findings are contested: some studies report a "compression phase" late in training (decreasing $I(T;X)$), others find no such phase once estimator bias (Problem 1) is controlled for. *Open*: does the IB curve, computed with provably consistent estimators, exhibit a universal compression phase across architectures and activation functions, or is the phenomenon an artefact of estimator bias interacting with saturating nonlinearities (e.g. tanh)? A rigorous resolution requires combining Problem 1's estimation theory with the optimisation dynamics of stochastic gradient descent.

3. Directed information and the capacity of channels with feedback. For channels without feedback, capacity is $C = \max_{p(x)} I(X;Y)$ (Part 4.2). When the encoder may use feedback (each X_i may depend on $Y_1, \dots, Y_{i-1}$), mutual information $I(X^n;Y^n)$ is no longer the right quantity — it must be replaced by Massey's *directed information* $I(X^n \to Y^n) = \sum_{i=1}^{n} I(X^i; Y_i | Y^{i-1})$, which respects the causal/temporal structure (Massey 1990; see also Pearl's causal calculus, cited in the Bibliography). For many channels with feedback, capacity is unchanged from the feedback-free case (e.g. DMCs, by Shannon 1956), but for channels with memory the feedback capacity is generally strictly larger and is known in closed form only for a handful of cases (the Gaussian channel with first-order autoregressive noise, via the Schalkwijk–Kailath scheme). *Open*: a general formula or efficiently computable characterisation of feedback capacity for finite-state channels with memory, and the corresponding directed-information analogue of the Blahut–Arimoto algorithm (Algorithm 5).

4. The capacity region of the general interference channel. Appendix E previewed the two-user Gaussian interference channel, whose capacity is known only "to within one bit" (Etkin–Tse–Wang 2008) via the Han–Kobayashi inner bound and a matching outer bound. For non-Gaussian interference channels, for channels with more than two users, or for the *exact* (not within-a-constant) capacity region even in the Gaussian two-user case, the problem remains open after more than five decades. The difficulty is fundamentally information-theoretic: unlike the MAC (Part E, solved via a single auxiliary "common message" argu-

ment), the interference channel's optimal coding strategy may require splitting each user's message into private and common parts in a way whose optimal split is itself unknown in general. *Open*: either an exact characterisation of the two-user Gaussian interference channel capacity region, or a proof that the Han–Kobayashi region is exactly optimal (matching converse).

5. Individual-sequence information theory and Kolmogorov complexity. The AEP (Part 3.1) and the source coding theorem are statements about *ensembles* of sequences (probabilistic averages over a source p). Kolmogorov complexity $K(x)$ — the length of the shortest program producing a specific string x — offers an alternative, ensemble-free notion of the "information content" of an individual sequence, with $\mathbb{E}_p[K(X^n)] = nH(\mathcal{X}) + O(\log n)$ for computable p (Li–Vitányi, cited in the Bibliography), but $K(x)$ itself is uncomputable. Universal source codes (LZ78, Algorithm 7; PPM; context-tree weighting) approximate $K(x)$ without computing it, achieving redundancy $O(\log n/n)$ for stationary ergodic sources. *Open*: for sources that are *not* stationary (e.g. piecewise-stationary, or governed by a slowly time-varying parameter, as in non-stationary network traffic or financial time series), what is the minimax redundancy of universal coding, and which algorithms — if any — achieve it adaptively without knowledge of the (unknown, time-varying) number of "regimes"?

6. Quantum extensions of the entropy and channel-capacity framework. Replacing probability distributions with density matrices ρ on a Hilbert space, the von Neumann entropy $S(\rho) = -(\rho \log \rho)$ generalises $H(X)$, and quantum mutual information, relative entropy, and channel capacities (Holevo capacity, quantum capacity for transmitting qubits) generalise Parts 2 and 4. Many classical results carry over (DPI, Pinsker-type inequalities, a quantum AEP), but *additivity* questions — whether the capacity of two channels used jointly equals the sum of their individual capacities — fail in the quantum setting for several capacity notions, a phenomenon with no classical analogue (superadditivity of quantum channel capacity under entanglement-assisted coding). *Open*: a complete classification of which quantum capacity notions are additive, and tight bounds on the maximal superadditivity gap as a function of channel dimension — directly relevant to the design of quantum communication networks and to bounds on classical simulation of quantum channels.

7. Second-order coding rates beyond the point-to-point channel. The strong converse and channel dispersion V (Part 4.5, Polyanskiy–Poor–Verdú) give a precise $nC - \sqrt{nV}Q^{-1}(\epsilon)$ characterisation of finite-blocklength performance for *point-to-point* DMCs. For multi-user settings (MAC, BC, interference channel of Problem 4), even the *first-order* capacity region may be open (Problem 4), let alone second-order refinements. Where first-order capacity *is* known (e.g. the MAC), the second-order region — how quickly the achievable rate region "shrinks" towards the boundary as a function of n and target error probability ϵ, and in which *direction* in the (R_1, R_2) plane — is known only for special cases (e.g. MACs with a single dominant user). *Open*: a general second-order theory for the MAC capacity region, including the geometry of the dispersion-dependent "rounding" of the corner points of the first-order region.

8. Maximum entropy under non-linear and structural constraints. The maximum-entropy duality of Appendix B is exact for *linear* (moment) constraints, yielding exponential-family solutions. Many modern applications — maximum-entropy models of network topology with given degree *sequences* (not just average degree), or maximum-entropy priors respecting a known causal graph structure (Pearl, cited in the Bibliography) — impose constraints that are combinatorial or structural rather than linear functionals of p. The resulting optimisation problems are generally non-convex in the natural parameterisation, though they may admit convex reformulations via lifting (as Bregman-divergence theory, Appendix B, suggests). *Open*: a general theory of maximum-entropy inference under graph-structural con-

straints, including efficient algorithms with convergence guarantees analogous to Blahut–Arimoto (Algorithm 5) — this is the entry point to the information-geometric perspective developed in Volume III of this series, where the exponential-family / Bregman-divergence duality of Appendix B is given its full Riemannian-geometric treatment.

A Measure-Theoretic Foundations of Entropy

Why measure theory? The definitions in the main text handle discrete and continuous random variables as separate cases. Measure theory unifies them and handles singular distributions (e.g., Cantor distributions) that are neither discrete nor continuous.

General Entropy (Measure-Theoretic)

Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space and $X : \Omega \to \mathcal{X}$ a measurable map, inducing measure μ_X on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$. If μ_X is dominated by a σ-finite reference measure ν (i.e., $\mu_X \ll \nu$), with Radon–Nikodym derivative $f = d\mu_X/d\nu$, then the entropy of X relative to ν is

$$H_\nu(X) = -\int_{\mathcal{X}} f(x) \log f(x)\, d\nu(x) = -\mathbb{E}[\log f(X)].$$

The discrete case uses $\nu =$ counting measure; the continuous case uses $\nu =$ Lebesgue measure.

KL divergence in general. For probability measures P and Q on $(\mathcal{X}, \mathcal{B})$:

$$D_{\mathrm{KL}}(P \,\|\, Q) = \begin{cases} \int \log \frac{dP}{dQ}\, dP & \text{if } P \ll Q \\ +\infty & \text{otherwise.} \end{cases}$$

This is the most general form of KL divergence, valid for any pair of probability measures.

Mutual information as KL divergence. $I(X;Y) = D_{\mathrm{KL}}(\mu_{XY} \,\|\, \mu_X \otimes \mu_Y)$, where μ_{XY} is the joint measure and $\mu_X \otimes \mu_Y$ is the product measure. This form is valid whether X, Y are discrete, continuous, or mixed.

Conditional entropy in measure theory. The conditional entropy $H(X|Y)$ is defined via the disintegration theorem: if the regular conditional distribution $\mu_{X|Y=y}$ exists, then

$$H(X|Y) = \int H_\nu(X|Y=y)\, d\mu_Y(y) = \mathbb{E}_Y[H_\nu(X|Y)].$$

Entropy of a mixed distribution. Let X have distribution $P = \lambda P_d + (1-\lambda)P_c$ where P_d is discrete (mass on countable set S) and P_c is absolutely continuous with density f_c. Then P has no density with respect to Lebesgue measure (P_d is singular), so differential entropy is undefined. However, $I(X;Y)$ remains well-defined via the general form.

Example — Entropy of a zero-inflated exponential

X is zero with probability λ and $\mathrm{Exp}(\mu)$ with probability $1-\lambda$. Distribution: $P = \lambda\delta_0 + (1-\lambda)\mathrm{Exp}(\mu)$. The natural reference measure $\nu = \delta_0 + \mathrm{Leb}$ gives density $f(0) = \lambda$ (with respect to δ_0) and $f(x) = (1-\lambda)\mu e^{-\mu x}$ for $x > 0$.

$$\begin{aligned} H_\nu(X) &= -\lambda \log \lambda - \int_0^\infty (1-\lambda)\mu e^{-\mu x} \log[(1-\lambda)\mu e^{-\mu x}]\, dx \\ &= -\lambda \log \lambda - (1-\lambda)\,[\log((1-\lambda)\mu) - 1] \\ &= h(\lambda) + (1-\lambda)[1 - \log \mu] \text{ nats,} \end{aligned}$$

where $h(\lambda) = -\lambda \log \lambda - (1-\lambda)\log(1-\lambda)$. For $\lambda = 0.3$, $\mu = 2$: $h(0.3) = 0.611$ nats, $(1-0.3)(1-\ln 2) = 0.7 \times 0.307 = 0.215$. $H_\nu(X) = 0.611 + 0.215 = 0.826$ nats.

B Convex Duality, Legendre Transform, and Entropy

The Legendre–Fenchel transform. For a convex function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, its convex conjugate is

$$f^*(\boldsymbol{\lambda}) = \sup_{\boldsymbol{\theta} \in \mathbb{R}^n} \{\boldsymbol{\lambda}^\top \boldsymbol{\theta} - f(\boldsymbol{\theta})\}.$$

Log-partition function and entropy. Let $\mathcal{X} = \{1, \dots, k\}$ and define the log-partition function $A(\boldsymbol{\theta}) = \log \sum_{x=1}^{k} e^{\theta_x}$ (a convex function of $\boldsymbol{\theta}$). Its conjugate is

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta}\{\boldsymbol{\mu}^\top \boldsymbol{\theta} - A(\boldsymbol{\theta})\} = \begin{cases} -H(\boldsymbol{\mu}) & \text{if } \boldsymbol{\mu} \in \Delta_k \\ +\infty & \text{otherwise.} \end{cases}}$$

where $\Delta_k = \{\boldsymbol{\mu} \geq 0 : \sum_i \mu_i = 1\}$ is the probability simplex. Thus *entropy is the negative conjugate of the log-partition function*: the convex duality between exponential families and entropy is exact.

Maximum entropy as a convex optimisation. The maximum entropy problem

$$\max_{\boldsymbol{\mu} \in \Delta_k} H(\boldsymbol{\mu}) \quad \text{s.t.} \quad A\boldsymbol{\mu} = \mathbf{b}$$

(for a matrix A encoding moment constraints) has dual

$$\min_{\boldsymbol{\lambda}} \{A^*(-\boldsymbol{\lambda}) + \mathbf{b}^\top \boldsymbol{\lambda}\} = \min_{\boldsymbol{\lambda}\{H^*(\boldsymbol{\mu}^*) + \boldsymbol{\lambda}^\top (\mathbf{b} - A\boldsymbol{\mu}^*)\},}$$

and the solution is always an exponential family distribution: $\mu^*(x) \propto \exp(\boldsymbol{\lambda}^{*\top} A_x)$.

Example — Maximum entropy with mean constraint

$\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$, constraint: $\mathbb{E}[X] = 4$ (biased die). Exponential family: $p^*(x) \propto e^{\lambda x}$, with λ chosen to satisfy $\sum_x x e^{\lambda x} / \sum_x e^{\lambda x} = 4$. Numerically: $\lambda \approx 0.284$. $p^*(x) \propto (1, 1.328, 1.764, 2.343, 3.111, 4.132)$, normalised: $p^* \approx (0.072, 0.096, 0.127, 0.169, 0.224, 0.298)$ (NB: $\sum p^* = 0.986 \approx 1$ after rounding). Entropy: $H(p^*) \approx 2.44$ bits $< \log_2 6 = 2.585$ bits; the mean constraint reduces entropy from the unconstrained maximum.

Bregman divergences and generalised entropy. Given a strictly convex differentiable function ϕ, the *Bregman divergence* is

$$B_\phi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \phi(\boldsymbol{\mu}) - \phi(\boldsymbol{\nu}) - \langle \nabla\phi(\boldsymbol{\nu}), \boldsymbol{\mu} - \boldsymbol{\nu} \rangle \geq 0.$$

KL divergence is the Bregman divergence associated with negative entropy $\phi(\boldsymbol{\mu}) = -H(\boldsymbol{\mu})$: $B_\phi(\mathbf{p}, \mathbf{q}) = D_{\mathrm{KL}}(\mathbf{p} \,\|\, \mathbf{q})$. The squared Euclidean distance corresponds to $\phi(\boldsymbol{\mu}) = \|\boldsymbol{\mu}\|_2^2/2$. The entire theory of exponential families and maximum entropy can be developed in terms of Bregman divergences.

C Entropy Bounds in Combinatorics

Shearer's lemma. Let $X_1, \dots, X_n$ be discrete random variables and $\mathcal{S} \subseteq 2^{[n]}$ a collection of subsets of $[n]$ such that each $i \in [n]$ belongs to at least t subsets in $\mathcal{S}$. Then

$$t \cdot H(X_1, \dots, X_n) \leq \sum_{S \in \mathcal{S}} H(X_S),$$

where $X_S = (X_i)_{i \in S}$.

Counting via entropy. Shearer's lemma can be used to prove combinatorial inequalities. For instance, let $\mathcal{F}$ be a family of subsets of $[n]$, each of size k. Choosing a uniform random $F \in \mathcal{F}$ and $i \in [n]$ with $i \in F$ (uniform over containments), Shearer gives:

$$|\mathcal{F}| \leq \binom{n}{k}^{1/k},$$

recovering the Loomis–Whitney inequality as a special case.

Entropy method for graph theory. For a graph G on n vertices with m edges, let X be a uniform random vertex. Then $H(X) = \log n$. Consider the pair (X, Y) where Y is a uniform random neighbour of X: $I(X;Y) = \log n - H(X|Y) \leq \log n$. This gives (after more work) the *entropy proof of the friendship paradox*: the average degree of a neighbour of a uniform random vertex is at least the average degree of the vertex itself.

Example — Entropy bound on error-correcting codes

An (n, k, d) binary linear code has 2^k codewords of length n with minimum Hamming distance d. The Singleton bound: $k \leq n - d + 1$. Entropy proof: Let C be a uniform random codeword. $H(C) = k$ bits. Any $d-1$ positions are determined by the remaining $n-d+1$ positions (since any $n-d+1$ positions determine the codeword uniquely when $d \geq 2$). Thus by the chain rule and DPI: $k = H(C) \leq H(C_{[n-d+1]}) \leq n - d + 1$. Numerical: a Reed–Solomon code over GF(2^8) with $n = 255$, $d = 33$ satisfies $k \leq 255 - 33 + 1 = 223$ coded symbols (each 8 bits). The RS(255,223) code achieves this bound and is used in CD/DVD error correction.

Hamming bound (sphere-packing). An (n, M, d) binary code (not necessarily linear) satisfies:

$$M \leq \frac{2^n}{\sum_{i=0}^{\lfloor (d-1)/2 \rfloor} \binom{n}{i}}.$$

Entropy proof: the Hamming balls of radius $t = \lfloor (d-1)/2 \rfloor$ around the M codewords are disjoint (by the distance constraint) and contained in $\{0,1\}^n$. Counting: $M \cdot |\text{Ball}(t)| \leq 2^n$. For large n with $d = \delta n$: $|\text{Ball}(\delta n/2)| \approx 2^{nh(\delta/2)}$, giving $\frac{\log M}{n} \leq 1 - h(\delta/2)$ (Hamming bound).

D Entropy and Hypothesis Testing: Sanov's Theorem

Sanov's Theorem

Let $X_1, \ldots, X_n$ be i.i.d. with distribution q on $\mathcal{X}$, and let $\hat{P}_n$ denote the empirical distribution. For any convex set $\Gamma \subseteq \mathcal{M}_1(\mathcal{X})$ of probability distributions:

$$\Pr(\hat{P}_n \in \Gamma) \doteq e^{-n \inf_{p \in \Gamma} D_{\mathrm{KL}}(p \,\|\, q)},$$

in the sense that $\frac{1}{n} \log \Pr(\hat{P}_n \in \Gamma) \to -\inf_{p \in \Gamma} D_{\mathrm{KL}}(p \,\|\, q)$.

Interpretation. The probability that the empirical distribution falls in a set Γ (even though the true distribution $q \notin \Gamma$) decays exponentially at rate equal to the minimum KL divergence from q to any distribution in Γ. This is the *information-theoretic large deviation principle*: the most likely atypical empirical distribution is the one closest (in KL) to q.

Proof sketch. The probability that $\hat{P}_n = P$ (for rational P with denominator n) is exactly $q^n(\mathcal{T}(P)) = e^{-n(D_{\mathrm{KL}}(P\,\|\,q)+o(1))}$ by the method of types. The set Γ is a union of type classes; the dominant contribution as $n \to \infty$ comes from the type $P^* \in \Gamma$ minimising $D_{\mathrm{KL}}(P\,\|\,q)$.

Example — Sanov's theorem for a biased coin

True distribution $q = (0.5, 0.5)$ (fair coin). Hypothesis: $\Gamma = \{p : p(0) \le 0.3\}$. $\inf_{p\in\Gamma} D_{\mathrm{KL}}(p\,\|\,q) = D_{\mathrm{KL}}((0.3,0.7)\,\|\,(0.5,0.5)) = 0.3\log(0.6) + 0.7\log(1.4) = 0.3(-0.737) + 0.7(0.336) = -0.221 + 0.235 = 0.099$ nats. After $n = 100$ flips: $\Pr(\hat{p}(0) \le 0.3) \approx e^{-100\times 0.099} = e^{-9.9} \approx 5 \times 10^{-5}$. Exact binomial: $\Pr(\hat{p}(0) \le 30/100) = \Pr(\mathrm{Bin}(100, 0.5) \le 30) \approx 3.9 \times 10^{-5}$. Sanov's approximation is accurate to a factor of ≈ 1.3.

Chernoff bound as a special case of Sanov. For $X_i \in \{0,1\}$ with $q = \mathrm{Bernoulli}(p)$, Sanov's theorem with $\Gamma = \{r : r \ge a\}$ (for $a > p$) gives:

$$\Pr\Big(\frac{1}{n}\sum X_i \ge a\Big) \doteq e^{-nD_{\mathrm{KL}}(a\,\|\,p)},$$

where $D_{\mathrm{KL}}(a\,\|\,p) = a\log(a/p) + (1-a)\log((1-a)/(1-p))$. This is the Chernoff bound in its tightest form.

Minimax hypothesis testing. For testing $H_0 : q$ vs $H_1 : p$ with n i.i.d. observations, the minimum *Bayes error* under equal prior is:

$$P_e^* = \frac{1}{2} - \frac{1}{2}\|\hat{P}_n - Q_n\|_{\mathrm{TV}} \doteq \frac{1}{2}e^{-nC(p,q)},$$

where $C(p,q) = -\min_{\lambda\in[0,1]} \log \sum_x p(x)^\lambda q(x)^{1-\lambda}$ is the Chernoff information (the best achievable error exponent). By Pinsker's inequality, $C(p,q) \ge \frac{1}{2}\min(D_{\mathrm{KL}}(p\,\|\,q)\,, D_{\mathrm{KL}}(q\,\|\,p))$.

E Preview of Network Information Theory

Multiple access channel (MAC). Two senders X_1, X_2 transmit to one receiver Y through $p(y|x_1,x_2)$. The capacity region $\mathcal{C}_{\mathrm{MAC}}$ is the convex hull of all rate pairs (R_1, R_2) with

$$\begin{aligned} R_1 &\le I(X_1; Y | X_2), \\ R_2 &\le I(X_2; Y | X_1), \\ R_1 + R_2 &\le I(X_1, X_2; Y), \end{aligned}$$

for some $p(x_1)p(x_2)$ (independent inputs). This is Shannon's MAC theorem (Ahlswede, Liao, 1972).

Broadcast channel (BC). One sender transmits to two receivers. The capacity region is known only for degraded BCs (Bergmans–Cover–Thomas): superposition coding achieves the boundary.

Interference channel. Two sender–receiver pairs sharing the medium. The capacity region is a longstanding open problem; inner bounds (Han–Kobayashi 1981) and outer bounds (Etkin–Tse–Wang 2008) are known. The Gaussian interference channel's capacity to within 1 bit was established by Etkin–Tse–Wang using lattice codes and the Z-channel approximation.

Relay channel. A sender communicates to a receiver via a relay that forwards information. The capacity is unknown in general; the decode-and-forward and compress-and-forward

strategies (Cover–El Gamal 1979) give achievable rates. For the degraded relay channel, capacity equals:

$$C = \max_{p(x,x_1)} \min\{I(X, X_1; Y), I(X; Y, Y_1|X_1)\}.$$

Secret key capacity. For a source $p(x, y)$ observed by Alice (X^n) and Bob (Y^n) and an eavesdropper (Z^n correlated with X^n, Y^n), the secret key capacity is

$$C_{\text{key}} = I(X;Y) - I(X;Z) = I(X;Y|Z).$$

This is the Maurer–Wolf theorem (1993), establishing that common randomness minus the eavesdropper's information is the fundamental limit for secret key agreement.

Example — Secret key capacity for a Gaussian source

(X, Y, Z) jointly Gaussian: $X \sim \mathcal{N}(0,1)$, $Y = X + N_1$ ($N_1 \sim \mathcal{N}(0, \sigma_1^2)$), $Z = X + N_2$ ($N_2 \sim \mathcal{N}(0, \sigma_2^2)$), independent noise.

$$I(X;Y) = \frac{1}{2}\log_2\left(1 + \frac{1}{\sigma_1^2}\right),$$
$$I(X;Z) = \frac{1}{2}\log_2\left(1 + \frac{1}{\sigma_2^2}\right),$$
$$C_{\text{key}} = \frac{1}{2}\log_2\left(\frac{1 + 1/\sigma_1^2}{1 + 1/\sigma_2^2}\right) = \frac{1}{2}\log_2\left(\frac{\sigma_2^2(1+\sigma_1^2)}{\sigma_1^2(1+\sigma_2^2)}\right).$$

For $\sigma_1 = 0.5$ (good Bob channel) and $\sigma_2 = 2$ (poor Eve channel): $C_{\text{key}} = \frac{1}{2}\log_2(\frac{4\times 1.25}{0.25\times 5}) = \frac{1}{2}\log_2(4) = 1$. One bit of secret key per channel use can be distilled from Bob's advantage over Eve.

Rate-distortion theory (preview for Vol. II). When lossless compression is impossible or unnecessary, one allows a distortion $d(x, \hat{x})$ between source symbol x and reconstruction $\hat{x}$. The rate-distortion function

$$R(D) = \min_{p(\hat{x}|x):\, \mathbb{E}[d(X,\hat{X})] \le D} I(X; \hat{X})$$

is the minimum rate for compression at distortion level D. For a Gaussian source with MSE distortion $d(x, \hat{x}) = (x - \hat{x})^2$ and power P: $R(D) = \frac{1}{2}\log_2(P/D)$ for $0 \le D \le P$ (and 0 otherwise). This is Shannon's reverse water-filling formula, connecting lossy compression to the Gaussian channel capacity.

F Proofs of Selected Theorems

This appendix collects fuller proofs of results that were stated, but only sketched, in the main text — for readers who want to see every step of the argument in one place.

F.1 Concavity of entropy.

Theorem

$H(\mathbf{p}) = -\sum_{i=1}^{k} p_i \log p_i$ is a concave function of $\mathbf{p} \in \Delta_k$.

Proof. The function $\phi(t) = -t \log t$ (with $\phi(0) = 0$) has second derivative $\phi''(t) = -1/t < 0$ for $t > 0$, so ϕ is strictly concave on $(0, 1]$. $H(\mathbf{p}) = \sum_i \phi(p_i)$ is a sum of concave functions of the individual coordinates p_i, and a sum of concave functions (each depending on a different coordinate) composed with a linear map (here, the identity on coordinates) is concave on the domain where the linear constraints $\sum_i p_i = 1, p_i \geq 0$ hold. More directly: for $\mathbf{p}, \mathbf{q} \in \Delta_k$ and $\lambda \in [0, 1]$, $H(\lambda\mathbf{p} + (1-\lambda)\mathbf{q}) = \sum_i \phi(\lambda p_i + (1-\lambda) q_i) \geq \sum_i [\lambda\phi(p_i) + (1-\lambda)\phi(q_i)] = \lambda H(\mathbf{p}) + (1-\lambda) H(\mathbf{q})$, where the inequality applies ϕ's concavity coordinatewise.

Consequence. Concavity of H is what makes the maximum-entropy problem (Appendix B) a well-posed concave maximisation with a unique maximiser whenever the feasible set $\Delta_k \cap \{A\mathbf{p} = \mathbf{b}\}$ is non-empty and convex (it is, being an intersection of the simplex with affine constraints).

F.2 Subadditivity of joint entropy.

Theorem

$H(X_1, \dots, X_n) \leq \sum_{i=1}^{n} H(X_i)$, with equality iff the X_i are mutually independent.

Proof by induction. For $n = 2$: $H(X_1, X_2) = H(X_1) + H(X_2|X_1) \leq H(X_1) + H(X_2)$ since $H(X_2|X_1) \leq H(X_2)$ (Problem 1.3, via the log-sum inequality / non-negativity of $I(X_1; X_2)$). For the inductive step, assume the claim for $n-1$ variables. By the chain rule, $H(X_1, \dots, X_n) = H(X_1, \dots, X_{n-1}) + H(X_n|X_1, \dots, X_{n-1}) \leq H(X_1, \dots, X_{n-1}) + H(X_n)$ (again using that conditioning cannot increase entropy, now conditioning on the whole tuple $(X_1, \dots, X_{n-1})$), and by the inductive hypothesis $H(X_1, \dots, X_{n-1}) \leq \sum_{i=1}^{n-1} H(X_i)$. Combining: $H(X_1, \dots, X_n) \leq \sum_{i=1}^{n-1} H(X_i) + H(X_n) = \sum_{i=1}^{n} H(X_i)$. Equality propagates backward through each step only if every conditional entropy term equals its unconditional counterpart, i.e. $X_n \perp (X_1, \dots, X_{n-1})$ *and* (by the inductive hypothesis's equality condition) $X_1, \dots, X_{n-1}$ are mutually independent — together these give full mutual independence of $X_1, \dots, X_n$.

F.3 The asymptotic equipartition property (full statement and proof).

Theorem (AEP)

If $X_1, X_2, \dots$ are i.i.d. with $H(X) < \infty$, then $-\frac{1}{n} \log p(X_1, \dots, X_n) \to H(X)$ in probability as $n \to \infty$.

Proof. Since X_i are i.i.d., $\log p(X_1, \dots, X_n) = \sum_{i=1}^{n} \log p(X_i)$, a sum of i.i.d. random variables $Y_i = \log p(X_i)$ with $\mathbb{E}[Y_i] = \sum_x p(x) \log p(x) = -H(X)$ (assumed finite). By the weak law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} Y_i \to \mathbb{E}[Y_1] = -H(X)$ in probability, i.e. $-\frac{1}{n}\sum_{i=1}^{n} \log p(X_i) = -\frac{1}{n} \log p(X_1, \dots, X_n) \to H(X)$ in probability.

Remark. The "in probability" mode of convergence is precisely what justifies the definition of the typical set $A_\varepsilon^{(n)}$ (Part 3): for any $\varepsilon > 0$ and $\delta > 0$, there exists n_0 such that for $n \geq n_0$,

$\Pr(X^n \in A_\varepsilon^{(n)}) > 1 - \delta$. A stronger mode of convergence (almost sure, via the strong law of large numbers, given a second-moment condition $\mathrm{Var}[\log p(X_1)] < \infty$) yields the Shannon–McMillan–Breiman theorem for general stationary ergodic sources, of which the i.i.d. case is the simplest instance.

F.4 Kraft–McMillan theorem (both directions).

Theorem

A sequence of positive integers $\ell(1), \dots, \ell(k)$ are the codeword lengths of *some* binary prefix code on k symbols if and only if $\sum_{i=1}^{k} 2^{-\ell(i)} \leq 1$.

Proof ($\Rightarrow$, *prefix code exists* $\Rightarrow$ *Kraft holds*). Represent the binary prefix code as a binary tree where each codeword is a leaf at depth $\ell(i)$. At depth $\ell_{\max} = \max_i \ell(i)$, the full binary tree has $2^{\ell_{\max}}$ nodes. Each codeword of length $\ell(i)$, being an internal-node-free leaf, "occupies" (i.e. has as descendants, none of which can be other codewords by the prefix property) exactly $2^{\ell_{\max} - \ell(i)}$ nodes at depth $\ell_{\max}$, and these sets of descendants are disjoint across different codewords (again by the prefix property — no codeword is an ancestor of another). Summing: $\sum_i 2^{\ell_{\max} - \ell(i)} \leq 2^{\ell_{\max}}$, i.e. $\sum_i 2^{-\ell(i)} \leq 1$.

Proof ($\Leftarrow$, *Kraft holds* $\Rightarrow$ *prefix code exists*). Order the lengths $\ell(1) \leq \ell(2) \leq \cdots \leq \ell(k)$. Construct codewords greedily: assign to symbol i the binary string corresponding to $\sum_{j<i} 2^{-\ell(j)}$, truncated/padded to $\ell(i)$ bits. One verifies this assignment is a valid prefix code precisely because the partial sums $\sum_{j \leq i} 2^{-\ell(j)} \leq 1$ (by Kraft's hypothesis) ensure no overlap between the "intervals" $[\sum_{j<i} 2^{-\ell(j)}, \sum_{j<i} 2^{-\ell(j)} + 2^{-\ell(i)})$ assigned to distinct symbols — this is the same construction underlying Shannon–Fano coding and the foundational link to arithmetic coding (Vol. II).

F.5 Convexity of mutual information in the channel (for fixed input).

Theorem

For fixed $p(x)$, $I(X;Y)$ is a convex function of the channel $p(y|x)$.

Proof sketch. Write $I(X;Y) = \sum_x p(x) D_{\mathrm{KL}}(p(\cdot|x) \,\|\, p(\cdot))$ where $p(y) = \sum_{x'} p(x')p(y|x')$ depends linearly on the channel. KL divergence $D_{\mathrm{KL}}(P \,\|\, Q)$ is jointly convex in (P, Q) (a standard fact, itself following from the log-sum inequality applied pointwise and summed). Here $P = p(\cdot|x)$ is one "block" of the channel (varying directly), and $Q = p(\cdot)$ is a *linear* function of the whole channel (an average over x' with fixed weights $p(x')$). The composition of a jointly convex function with an affine map (in the second argument) and identity (in the first) remains convex in the channel as a whole; summing over x with non-negative weights $p(x)$ preserves convexity. This convexity (in the channel, for fixed input) and the concavity in the input (for fixed channel, proven via similar KL-convexity arguments applied the other way) together justify the alternating-maximisation structure of the Blahut–Arimoto algorithm: each sub-problem (optimise $p(x)$ for fixed auxiliary Q, or optimise Q for fixed $p(x)$) is a well-posed concave maximisation with a unique solution characterisable via Lagrange multipliers / Bayes' rule respectively.

F.6 Shannon's channel coding theorem — achievability sketch (random coding).

Theorem (Achievability)

For any $R < C = \max_{p(x)} I(X;Y)$, there exists a sequence of $(2^{nR}, n)$ codes with maximum error probability $\to 0$ as $n \to \infty$.

Proof sketch (*random coding + joint typicality decoding*). Fix $p(x)$ achieving C. Generate 2^{nR} codewords $X^n(m)$, $m = 1, \dots, 2^{nR}$, each i.i.d. $\sim \prod_i p(x_i)$ (a "random codebook"). To decode received Y^n, the receiver finds the unique m such that $(X^n(m), Y^n)$ are *jointly typical* (i.e. $(X^n(m), Y^n) \in A_\varepsilon^{(n)}$ for the joint distribution $p(x, y)$). Two error events: (a) the true codeword $X^n(m^*)$ and Y^n are *not* jointly typical — by the joint AEP this probability $\to 0$; (b) some *other* codeword $X^n(m')$, $m' \neq m^*$, happens to be jointly typical with Y^n — since $X^n(m') \perp Y^n$ (independent random codewords), $\Pr((X^n(m'), Y^n) \in A_\varepsilon^{(n)}) \approx 2^{-nI(X;Y)} = 2^{-nC}$ (joint AEP for independent sequences). Union bound over the $2^{nR}-1$ other codewords: $\Pr(\text{error type (b)}) \lesssim 2^{nR} \times 2^{-nC} = 2^{-n(C-R)}$, which $\to 0$ whenever $R < C$. Averaging over random codebooks, the average error probability $\to 0$; hence *some* codebook achieves error $\to 0$ (and a standard expurgation argument removes the worst half of codewords to ensure *maximum*, not just average, error $\to 0$, at the cost of halving the rate — asymptotically negligible).

G Statistical and Reference Tables

G.1 Binary entropy function $h(p)$ for selected p.

p	$h(p)$ (bits)	p	$h(p)$ (bits)	p	$h(p)$ (bits)
0.01	0.0808	0.20	0.7219	0.50	1.0000
0.02	0.1414	0.25	0.8113	0.55	0.9928
0.05	0.2864	0.30	0.8813	0.60	0.9710
0.10	0.4690	0.35	0.9341	0.70	0.8813
0.15	0.6098	0.40	0.9710	0.80	0.7219

By symmetry, $h(p) = h(1-p)$, so values for $p > 0.5$ mirror those for $p < 0.5$.

G.2 BSC capacity $C = 1 - h(\epsilon)$ for selected crossover probabilities.

ϵ	$h(\epsilon)$	$C = 1 - h(\epsilon)$ (bits)
0.001	0.0114	0.9886
0.01	0.0808	0.9192
0.05	0.2864	0.7136
0.10	0.4690	0.5310
0.15	0.6098	0.3902
0.20	0.7219	0.2781
0.30	0.8813	0.1187
0.50	1.0000	0.0000

At $\epsilon = 0.5$, the channel output is independent of the input ($C = 0$): a fair-coin BSC conveys no information whatsoever, the information-theoretic analogue of "pure noise".

G.3 Gaussian channel capacity $C = \frac{1}{2}\log_2(1 + \text{SNR})$ for selected SNR.

SNR (linear)	SNR (dB)	C (bits/use)	Comment
1	0.00	0.500	Unity SNR
3	4.77	1.000	$C = 1$
7	8.45	1.500	
15	11.76	2.000	$C = 2$ (Problem 4.2)
31	14.91	2.500	
63	17.99	3.000	$C = 3$
100	20.00	3.330	

Note the pattern: $C = k$ (an integer number of bits) corresponds to SNR $= 2^{2k} - 1$ (i.e. $1, 3, 15, 63, \ldots$ for $k = 0.5, 1, 1.5, 2, \ldots$ — more precisely SNR $= 4^k - 1$), each additional bit of capacity requiring the SNR to (roughly) quadruple — the well-known "6 dB per bit" rule of thumb (since $10\log_{10} 4 \approx 6.02$).

G.4 Common probability distributions and their entropies / differential entropies.

Distribution	Parameters	Entropy / Differential entropy
Bernoulli(p)	$p \in [0,1]$	$h(p)$ bits
Geometric(p) on $\{1,2,...\}$	$p \in (0,1]$	$h(p)/p$ bits
Uniform on $\{1,\dots,k\}$	$k \in \mathbb{N}$	$\log_2 k$ bits
Poisson(λ)	$\lambda > 0$	$\approx \frac{1}{2}\log_2(2\pi e\lambda)$ for large λ (Gaussian approx.)
Uniform on $[a,b]$	$a < b$	$\ln(b-a)$ nats
Gaussian $\mathcal{N}(\mu,\sigma^2)$	$\mu \in \mathbb{R}, \sigma > 0$	$\frac{1}{2}\ln(2\pi e\sigma^2)$ nats
Exponential(λ)	$\lambda > 0$	$1 - \ln\lambda$ nats
Laplace (scale b)	$b > 0$	$\ln(2eb)$ nats
Multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	$\boldsymbol{\Sigma} \succ 0$, n dims	$\frac{1}{2}\ln((2\pi e)^n\lvert\boldsymbol{\Sigma}\rvert)$ nats

H Notation Index

Symbol	Meaning
$\mathcal{X}, \mathcal{Y}, \mathcal{Z}$	Alphabets (finite or countable sets) for random variables X, Y, Z.
$p(x)$, $p_X(x)$	Probability mass / density function of X at x.
$p(y\|x)$	Conditional distribution / channel transition probability.
$\Pr(\cdot)$	Probability of an event.
$\mathbb{E}[\cdot]$, $\operatorname{Var}(\cdot)$, $\operatorname{Cov}(\cdot,\cdot)$	Expectation, variance, covariance.
$H(X)$	Shannon entropy of X (bits unless stated otherwise).
$H(X,Y)$	Joint entropy.
$H(Y\|X)$	Conditional entropy of Y given X.
$I(X;Y)$	Mutual information between X and Y.
$D_{\mathrm{KL}}(P \,\|\, Q)$	Kullback–Leibler divergence from P to Q.
$h(p)$	Binary entropy function.
$h(X)$	Differential entropy of continuous X.
nats, bits	Units: natural logarithm base e (nats) or base 2 (bits).
Δ_k	Probability simplex on k outcomes: $\{\mathbf{p} \geq 0 : \sum_i p_i = 1\}$.
$A_\varepsilon^{(n)}$	ε-typical set of length-n sequences.
$\ell(x)$	Codeword length assigned to symbol x.
L, L^*	Expected codeword length; optimal expected codeword length.
$\mathcal{N}(\mu, \sigma^2)$	Gaussian (normal) distribution with mean μ, variance σ^2.
C	Channel capacity.
ϵ	Crossover (BSC) or erasure (BEC) probability.
P, N	Signal power, noise power (Gaussian channel).
V	Channel dispersion.
$Q^{-1}(\cdot)$	Inverse Gaussian Q-function (tail probability).
$\mathbb{R}, \mathbb{N}, \mathbb{Z}$	Real numbers, natural numbers, integers.
$\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$	Bold lowercase/uppercase for vectors and matrices.
∇, ∇^2	Gradient, Hessian.
$\propto$	”Proportional to”.
$\doteq$	Equality of exponential rates (large-deviations notation).
$(\cdot)^+$	Positive part, $\max(\cdot, 0)$.
$\mathbb{1}[\cdot]$	Indicator function.
$X - Y - Z$	Markov chain notation: Z conditionally independent of X given Y.

I Quick Reference: Generalized Entropies, Divergences, and Operational Exponents

This appendix collects, in tabular form, the generalized entropy and divergence families introduced in Sections 1.6, 2.5, and 3.7, together with their defining formulas, limiting cases, and operational interpretations. As with the Formula Reference sections of the main text, the purpose is rapid lookup rather than derivation; all formulas are derived in full in the sections cited.

I.1 The Rényi entropy family $H_\alpha(X) = \frac{1}{1-\alpha}\log\sum_x p(x)^\alpha$ (Section 1.6).

α	Formula	Operational meaning
0	$H_0 = \log\lvert\mathrm{supp}(p)\rvert$	Hartley entropy: log of the number of possible outcomes, ignoring probabilities entirely.
$1/2$	$H_{1/2} = 2\log\sum_x \sqrt{p(x)}$	Related to the Bhattacharyya coefficient and the Hellinger distance (Section 2.5).
1	$H_1 = H(X) = -\sum_x p(x)\log p(x)$	Shannon entropy: expected code length / average surprise; governs source coding (Part 3).
2	$H_2 = -\log\sum_x p(x)^2$	Collision entropy: $-\log\Pr(X = X')$ for an independent copy X'; governs universal hashing and the leftover hash lemma.
∞	$H_\infty = -\log\max_x p(x)$	Min-entropy: governs the single-shot guessing probability $P_{\text{guess}} = 2^{-H_\infty}$; the operative quantity in cryptography.

Monotonicity: $H_0 \geq H_{1/2} \geq H_1 \geq H_2 \geq H_\infty$, with equality throughout iff p is uniform on its support.

I.2 The Tsallis entropy family $S_q(X) = \frac{1}{q-1}(1 - \sum_x p(x)^q)$ (Section 1.6).

Property	Formula	Remark
Limit $q \to 1$	$S_1(X) = H(X)$	Shannon entropy recovered by L'Hôpital's rule, as for H_α.
Pseudo-additivity	$S_q(X,Y) = S_q(X) + S_q(Y) + (1-q)S_q(X)S_q(Y)$ for $X \perp Y$	Reduces to ordinary additivity only at $q = 1$; superextensive for $q < 1$, subextensive for $q > 1$.
Maximizer	q-exponential $p(x) \propto [1 - (1-q)\beta\varepsilon(x)]_+^{1/(1-q)}$	Power-law equilibrium distribution under a generalized energy constraint; reduces to the Boltzmann–Gibbs exponential as $q \to 1$.

I.3 The f-divergence family $D_f(p\|q) = \sum_x q(x) f(p(x)/q(x))$ (Section 2.5).

Divergence	Generator $f(t)$	Resulting formula
KL divergence	$t\log t$	$D_{\mathrm{KL}}(p\,\|\,q)=\sum_x p(x)\log\frac{p(x)}{q(x)}$
Reverse KL	$-\log t$	$D_{\mathrm{KL}}(q\,\|\,p)=\sum_x q(x)\log\frac{q(x)}{p(x)}$
Total variation	$\frac{1}{2}\lvert t-1\rvert$	$\|p-q\|_{\mathrm{TV}}=\frac{1}{2}\sum_x\lvert p(x)-q(x)\rvert$
χ^2-divergence	$(t-1)^2$	$\chi^2(p,q)=\sum_x\frac{(p(x)-q(x))^2}{q(x)}$
Squared Hellinger	$(\sqrt{t}-1)^2$	$2H^2(p,q)=\sum_x(\sqrt{p(x)}-\sqrt{q(x)})^2$
Jensen–Shannon	$t\log\frac{2t}{t+1}+\log\frac{2}{t+1}$	$\mathrm{JS}(p,q)=\frac{1}{2}D_{\mathrm{KL}}(p\,\|\,m)+\frac{1}{2}D_{\mathrm{KL}}(q\,\|\,m)$, $m=\frac{1}{2}(p+q)$; bounded by $\log 2$

All members satisfy $D_f(p\|q)\geq 0$ with equality iff $p=q$ (Jensen's inequality applied to the convex f, $f(1)=0$), are jointly convex in (p,q), and satisfy the data-processing inequality $D_f(Tp\|Tq)\leq D_f(p\|q)$ for any Markov kernel T.

I.4 The Rényi divergence family $D_\alpha(p\|q)=\frac{1}{\alpha-1}\log\sum_x p(x)^\alpha q(x)^{1-\alpha}$ (Section 1.6) and the Chernoff information (Section 3.7).

Quantity	Formula	Operational role
$D_0(p\|q)$	0 if $p\ll q$, else ∞	Trivial endpoint.
$D_{1/2}(p\|q)$	$-2\log\sum_x\sqrt{p(x)q(x)}$	Bhattacharyya bound on TV / Hellinger distance.
$D_1(p\|q)$	$D_{\mathrm{KL}}(p\,\|\,q)$	Stein exponent: governs $\beta_n^*(\epsilon)\approx 2^{-nD_{\mathrm{KL}}(p\,\|\,q)}$ in asymmetric hypothesis testing.
$D_\infty(p\|q)$	$\log\max_x\frac{p(x)}{q(x)}$	Worst-case log-likelihood ratio; used in differential privacy.
Chernoff information	$C(p,q)=\max_{0\leq\lambda\leq 1}(1-\lambda)D_\lambda(p\|q)$	Governs the Bayes error exponent $P_e^{*(n)}\approx 2^{-nC(p,q)}$ in symmetric hypothesis testing.

Monotonicity in α: $D_\alpha(p\|q)$ is non-decreasing in $\alpha\in(0,\infty)$ and non-negative throughout, with $D_\alpha(p\|q)=0$ for all α iff $p=q$.

I.5 Worked numerical cross-reference. For the running pair $p=\mathrm{Bernoulli}(0.5)$, $q=\mathrm{Bernoulli}(0.7)$ used throughout Sections 1.6, 2.5, and 3.7 (base-2 logarithms):

Quantity	Value (bits)	Source
$D_{\mathrm{KL}}(p\,\|\,q)$	0.1258	Section 2.5 / 3.7
$D_{\mathrm{KL}}(q\,\|\,p)$	0.1184	Section 2.5 / 3.7
$\|p-q\|_{\mathrm{TV}}$	0.2000	Section 2.5
$\chi^2(p,q)$	0.1600	Section 2.5
$2H^2(p,q)$	0.0422	Section 2.5
$D_2(p\|q)$	0.2141	Section 1.6
$D_{1/2}(p\|q)$	0.0615	Section 1.6
$C(p,q)$ (Chernoff)	≥ 0.0925	Section 3.7 (lower bound from $\lambda=0.2$)

The ordering $D_{1/2}<D_{\mathrm{KL}}(q\,\|\,p)<D_{\mathrm{KL}}(p\,\|\,q)<D_2$ instantiates the general monotonicity

$D_\alpha \le D_\beta$ for $\alpha < \beta$ (Section 1.6), while the Chernoff information $C(p,q)$, being a λ-weighted Rényi quantity rather than a single D_α, does not sit at a fixed point of this ordering — it is bounded above by $\min(D_{\mathrm{KL}}(p\,\|\,q)\,, D_{\mathrm{KL}}(q\,\|\,p))$ (a general fact: the Bayes exponent cannot exceed either one-sided Stein exponent) and the numerical value above is consistent with $C(p,q) \le D_{\mathrm{KL}}(q\,\|\,p) = 0.1184$.

J A Catalogue of Exponential Families

Section 2.6 showed that maximum-entropy distributions subject to moment constraints are exponential families $p_\theta(x) = h(x)\exp(\theta^\top T(x) - A(\theta))$, with the log-partition function $A(\theta)$ encoding the mean (∇A), variance ($\nabla^2 A$), and KL-divergence (Bregman divergence of A) structure of the family. This appendix tabulates these quantities for the distributions most frequently encountered in this volume and its companions, in the single natural-parameter ($k = 1$) case where the bookkeeping is most transparent; the multivariate generalizations (multinomial, multivariate Gaussian) follow the same pattern with $\theta, T(x) \in \mathbb{R}^k$ and A a function of a vector argument, as in Appendix B.

J.1 One-parameter discrete families.

Distribution	θ	$T(x)$	$A(\theta)$	$\mu = A'(\theta)$	$A''(\theta) = \mathrm{Var}(T(X))$
Bernoulli(p), $x \in \{0,1\}$	$\log \frac{p}{1-p}$	x	$\log(1+e^\theta)$	$\frac{e^\theta}{1+e^\theta} = p$	$p(1-p)$
Geometric(p) on $\{0,1,2,\ldots\}$	$\log(1-p)$	x	$-\log(1-e^\theta)$	$\frac{e^\theta}{1-e^\theta} = \frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\log\lambda$	x	e^θ	$e^\theta = \lambda$	λ

Verification for the Bernoulli row. $A(\theta) = \log(1+e^\theta)$ is the log-partition function since $Z(\theta) = \sum_{x\in\{0,1\}} e^{\theta x} = 1 + e^\theta$. Then $A'(\theta) = \frac{e^\theta}{1+e^\theta}$, which, writing $\theta = \log\frac{p}{1-p}$ so $e^\theta = \frac{p}{1-p}$, gives $A'(\theta) = \frac{p/(1-p)}{1+p/(1-p)} = \frac{p}{(1-p)+p} = p$ — the mean parameter is the success probability, as expected. And $A''(\theta) = A'(\theta)(1-A'(\theta)) = p(1-p)$, the Bernoulli variance, confirming the general identity $A''(\theta) = \mathrm{Var}_\theta(T(X))$ from Section 2.6 in this simplest case.

Verification for the Poisson row. $Z(\theta) = \sum_{x=0}^\infty \frac{\lambda^x}{x!} e^{\theta x}$ with the base measure $h(x) = 1/x!$ absorbing the factorial; writing λe^θ in place of the rate inside the sum, $Z(\theta) = \sum_x \frac{(\lambda e^\theta)^x}{x!} e^{-\lambda}\cdot e^\lambda$ — more directly, parametrizing the Poisson(μ) family by $\theta = \log\mu$ gives $p_\theta(x) = \frac{1}{x!}e^{\theta x}e^{-e^\theta}$ (matching the standard Poisson pmf $\frac{\mu^x}{x!}e^{-\mu}$ with $\mu = e^\theta$), so $A(\theta) = e^\theta$ directly. Then $A'(\theta) = e^\theta = \mu$ and $A''(\theta) = e^\theta = \mu$: *the Poisson distribution is the unique family for which the mean equals the variance*, a fact often used as a dispersion diagnostic in count-data modeling, and here seen to be a direct consequence of $A(\theta) = e^\theta$ satisfying $A = A' = A''$.

J.2 One-parameter continuous families.

Distribution	θ	$T(x)$	$A(\theta)$	$\mu = A'(\theta)$	$A''(\theta)$
Exponential(λ), $x \geq 0$	$-\lambda$	x	$-\log(-\theta)$	$-1/\theta = 1/\lambda$	$1/\theta^2 = 1/\lambda^2$
Gaussian $\mathcal{N}(\mu, \sigma_0^2)$, σ_0 fixed	μ/σ_0^2	x	$\sigma_0^2\theta^2/2$	$\sigma_0^2\theta = \mu$	σ_0^2

The exponential family in natural form. Writing $p_\theta(x) = e^{\theta x - A(\theta)}$ for $x \geq 0$ with $A(\theta) = -\log(-\theta)$ ($\theta < 0$ for normalizability), $p_\theta(x) = e^{\theta x}\cdot(-\theta) = (-\theta)e^{\theta x}$. Setting $\lambda = -\theta > 0$ gives $p_\theta(x) = \lambda e^{-\lambda x}$, the familiar Exp($\lambda$) density — confirming the parametrization in the table. The mean $\mu = -1/\theta = 1/\lambda$ and variance $A''(\theta) = 1/\theta^2 = 1/\lambda^2$ match the standard exponential-distribution moments exactly, and since $A'(\theta) = -1/\theta$ as well, this family

satisfies $A''(\theta) = (A'(\theta))^2$, i.e. Var $= \mu^2$ (coefficient of variation 1) — a special algebraic coincidence of $A(\theta) = -\log(-\theta)$ not shared by the other families in this catalogue.

Example — KL divergence between two Poisson distributions via the Bregman identity

By Section 2.6, $D_{\mathrm{KL}}\big(p_{\theta_1} \,\|\, p_{\theta_2}\big) = B_A(\theta_2, \theta_1) = A(\theta_2) - A(\theta_1) - A'(\theta_1)(\theta_2 - \theta_1)$. For Poisson, $A(\theta) = e^\theta$, $A'(\theta) = e^\theta$. With $\theta_i = \log \mu_i$,

$$D_{\mathrm{KL}}(\mathrm{Poisson}(\mu_1) \,\|\, \mathrm{Poisson}(\mu_2)) = e^{\theta_2} - e^{\theta_1} - e^{\theta_1}(\theta_2 - \theta_1)$$

$$= \mu_2 - \mu_1 - \mu_1 \log \frac{\mu_2}{\mu_1} = \mu_2 - \mu_1 + \mu_1 \log \frac{\mu_1}{\mu_2}.$$

Numerical check, $\mu_1 = 2, \mu_2 = 3$ (natural logarithm, nats): $\mu_2 - \mu_1 = 1$, $\mu_1 \log(\mu_1/\mu_2) = 2\log(2/3) = 2(-0.4055) = -0.8109$. So $D_{\mathrm{KL}}(\mathrm{Poisson}(2) \,\|\, \mathrm{Poisson}(3)) = 1 - 0.8109 = 0.1891$ nats. **Direct verification** via the definition $\sum_{x \geq 0} p_1(x) \log \frac{p_1(x)}{p_2(x)}$ with $p_i(x) = \frac{\mu_i^x e^{-\mu_i}}{x!}$: $\log \frac{p_1(x)}{p_2(x)} = x \log \frac{\mu_1}{\mu_2} - (\mu_1 - \mu_2) = x\log(2/3) + 1 = -0.4055x + 1$. Taking the expectation under $p_1 = \mathrm{Poisson}(2)$ (so $\mathbb{E}_{p_1}[X] = \mu_1 = 2$): $\mathbb{E}_{p_1}[-0.4055X + 1] = -0.4055(2) + 1 = -0.8109 + 1 = 0.1891$ nats, matching the Bregman-divergence formula exactly — and obtained with no infinite sum, only the single moment $\mathbb{E}_{p_1}[X] = \mu_1$.

J.3 The two-parameter Gaussian family. When both μ and σ^2 are free, $\mathcal{N}(\mu, \sigma^2)$ is a two-parameter exponential family with $T(x) = (x, x^2)$ and natural parameters $\theta_1 = \mu/\sigma^2$, $\theta_2 = -1/(2\sigma^2)$ (Section 2.6's worked example). The log-partition function is

$$A(\boldsymbol{\theta}) = -\frac{\theta_1^2}{4\theta_2} - \frac{1}{2}\log(-2\theta_2),$$

with gradient $\nabla A = (\mu, \mu^2 + \sigma^2)$ — the mean parameters are $(\mathbb{E}[X], \mathbb{E}[X^2])$, not (μ, σ^2) directly, illustrating that the "mean parameter" coordinate system $\boldsymbol{\mu} = \nabla A(\boldsymbol{\theta})$ of Section 2.6 is the vector of *moments* of the sufficient statistics, which for the Gaussian are the first and second moments rather than the mean and variance themselves (though the two are in bijection via $\sigma^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$). The Hessian $\nabla^2 A(\boldsymbol{\theta})$ — the Fisher information matrix in natural coordinates — is, after this change of variables, the familiar 2×2 Fisher information matrix for (μ, σ^2) with entries $1/\sigma^2$, 0, and $1/(2\sigma^4)$ (diagonal, reflecting the classical fact that the sample mean and sample variance are asymptotically independent and efficient estimators of μ and σ^2 respectively).

J.4 Conjugate priors as a Bayesian reading of the same table. Every row of this catalogue also describes a conjugate-prior pair in Bayesian inference: if the likelihood $p(x|\theta)$ belongs to an exponential family with sufficient statistic $T(x)$, the conjugate prior on θ has the same exponential-family form with $T(x)$ replaced by θ itself (and the roles of natural and mean parameters interchanged between likelihood and prior). For example, the Beta distribution is conjugate to the Bernoulli/Binomial likelihood, and the Gamma distribution is conjugate to the Poisson and Exponential likelihoods — in each case, the posterior update amounts to adding the observed sufficient statistic $\sum_i T(x_i)$ to the prior's natural parameters, a fact that is most transparent precisely in the natural-parameter coordinates θ of this table, and considerably less transparent in the original $(p, \lambda, \mu, \sigma^2, \ldots)$ parametrizations. This Bayesian reading will recur in Volume III's discussion of the Fisher metric as the unique (up to scale) Riemannian metric invariant under sufficient statistics (Čencov's theorem), where the natural and mean parameter coordinate systems of this appendix become, respectively, the affine coordinates of the exponential (e-) and mixture (m-) connections on the statistical manifold.

Extended Case Study: Information-Theoretic Design of a Deep-Space Communication Link

This case study applies essentially every tool developed in Volume I — entropy, mutual information, the AEP, source coding, channel capacity, water-filling, and dispersion — to a single realistic engineering problem: designing the downlink from a deep-space probe to Earth, from the raw image sensor to the decoded bitstream.

G.1 Mission parameters. A probe orbiting a distant target carries a camera producing 12-bit greyscale images of 1024×1024 pixels. The downlink uses a 20 W transmitter, a parabolic antenna with gain $G_T = 50$ dBi, and the ground station has receive gain $G_R = 70$ dBi. The Earth–probe distance is $d = 8 \times 10^{11}$ m (roughly 5.3 AU), carrier frequency $f = 8.4$ GHz (X-band), and the system noise temperature is $T_{\text{sys}} = 25$ K.

G.2 Link budget and received SNR. Free-space path loss: $L_{\text{fs}} = \left(\frac{4\pi df}{c}\right)^2$. In dB: $L_{\text{fs}}[\text{dB}] = 20\log_{10}(4\pi df/c)$. With $c = 3 \times 10^8$ m/s: $4\pi df/c = 4\pi \times 8 \times 10^{11} \times 8.4 \times 10^9/(3 \times 10^8) = 4\pi \times 2.24 \times 10^{13} = 2.815 \times 10^{14}$. $L_{\text{fs}}[\text{dB}] = 20\log_{10}(2.815 \times 10^{14}) = 20 \times 14.4495 = 288.99$ dB.

Transmit power: $P_T[\text{dBW}] = 10\log_{10}(20) = 13.01$ dBW. Received power: $P_R[\text{dBW}] = P_T + G_T + G_R - L_{\text{fs}} = 13.01 + 50 + 70 - 288.99 = -155.98$ dBW.

Noise power spectral density: $N_0 = k_B T_{\text{sys}}$, where $k_B = 1.38 \times 10^{-23}$ J/K. $N_0 = 1.38 \times 10^{-23} \times 25 = 3.45 \times 10^{-22}$ W/Hz, i.e. $N_0[\text{dBW/Hz}] = 10\log_{10}(3.45 \times 10^{-22}) = -214.62$ dBW/Hz.

The ratio P_R/N_0 (in dB-Hz): $P_R - N_0 = -155.98 - (-214.62) = 58.64$ dB-Hz. This figure — the "received signal-to-noise spectral density ratio" — is the fundamental quantity from which all achievable rates and required bandwidths are derived.

G.3 Source entropy: characterising the image data. The 12-bit pixel values are not uniformly distributed. Empirically (a typical result for natural and astronomical images), the marginal pixel distribution is well-approximated by a Laplace distribution centred on the predicted value from neighbouring pixels (after a simple predictive transform), with scale parameter $b \approx 14$ (in 12-bit units, range 0–4095). The differential entropy of a Laplace(b) distribution is $\ln(2eb)$ nats. For $b = 14$: $h = \ln(2 \times 2.71828 \times 14) = \ln(76.11) = 4.3327$ nats $= 6.2502$ bits.

Since the underlying signal is discretised to 12 bits (quantisation step $\Delta = 1$), the discrete entropy of the prediction-residual is approximately $h + \log_2(1/\Delta) = 6.2502$ bits per pixel (for $\Delta = 1$, $\log_2(1/\Delta) = 0$). Compare to the "raw" entropy if pixels were i.i.d. uniform over $[0, 4095]$: $\log_2 4096 = 12$ bits per pixel. The predictive transform plus the residual's concentrated (Laplace, not uniform) distribution reduces the per-pixel entropy from 12 to roughly 6.25 bits — a compression factor of $\approx 1.92\times$ *before any entropy coder is applied*, simply by exploiting spatial correlation. This is the information-theoretic justification for predictive (DPCM-style) preprocessing in lossless image compression, developed fully in Volume II.

G.4 Source coding: expected codeword length. Applying the source coding theorem (Part 3), an optimal entropy coder (e.g. an adaptive arithmetic coder matched to the empirical Laplace residual distribution) achieves expected length $L^* \to H(\text{residual}) \approx 6.25$ bits/pixel as block length grows, with a Huffman code (Part 3) typically achieving $L_{\text{Huffman}} \approx 6.3$–$6.4$ bits/pixel (within the $H(X) \le L < H(X) + 1$ guarantee, but in practice close to $H(X)$ for smooth/unimodal distributions like the Laplace).

Total data per image: $1024 \times 1024 \times 6.30 \approx 6.609 \times 10^6$ bits ≈ 826 KB (compressed), versus $1024 \times 1024 \times 12 = 1.258 \times 10^7$ bits ≈ 1.573 MB (uncompressed) — a $1.90\times$ reduction, consistent with the entropy estimate.

G.5 Channel capacity: how much bandwidth is needed? The Shannon–Hartley formula relates capacity (bits/second) to bandwidth W (Hz) and received SNR: $C = W\log_2(1 + \text{SNR})$, where $\text{SNR} = P_R/(N_0 W)$. Substituting: $C = W\log_2\left(1 + \frac{P_R/N_0}{W}\right)$.

Suppose the mission requires transmitting one compressed image every $\tau = 600$ s (one image every 10 minutes — a realistic deep-space cadence). Required rate: $R = 6.609 \times 10^6/600 = 11{,}015$ bits/s ≈ 11.0 kbps.

We need $C \geq R$, i.e. $W\log_2\left(1 + \frac{10^{5.864}}{W}\right) \geq 11{,}015$, where $10^{5.864} = 7.31 \times 10^5$ (converting 58.64 dB-Hz to linear).

Try $W = 2000$ Hz: $\text{SNR} = 7.31 \times 10^5/2000 = 365.5$. $C = 2000 \times \log_2(366.5) = 2000 \times 8.518 = 17{,}036$ bits/s. This exceeds $R = 11{,}015$, so $W = 2000$ Hz is more than sufficient.

Try $W = 1000$ Hz: $\text{SNR} = 7.31 \times 10^5/1000 = 731$. $C = 1000 \times \log_2(732) = 1000 \times 9.516 = 9516$ bits/s $< R = 11{,}015$. Insufficient.

Try $W = 1300$ Hz: $\text{SNR} = 7.31 \times 10^5/1300 = 562.3$. $C = 1300 \times \log_2(563.3) = 1300 \times 9.138 = 11{,}879$ bits/s $> R$. Sufficient with margin.

So a bandwidth between roughly 1100 and 1300 Hz suffices — an extremely narrow channel by terrestrial standards, reflecting the very low received power at 5.3 AU. This illustrates the fundamental link-budget trade-off: capacity grows only *logarithmically* in SNR but *linearly* in bandwidth, so for power-starved (low-SNR) links, narrow bandwidth with high spectral efficiency (many bits/s/Hz, here ≈ 9.1) is preferred over wide bandwidth with low spectral efficiency.

G.6 Channel dispersion: finite-blocklength penalty for image transfer. At $W \approx 1300$ Hz and a transmission window of $\tau = 600$ s, the blocklength is $n \approx 2W\tau = 2 \times 1300 \times 600 = 1.56 \times 10^6$ channel uses (using the Nyquist sampling rate $2W$). The Gaussian channel dispersion is $V = \frac{1}{2}\log_2^2(e) \times \frac{\text{SNR}(\text{SNR}+2)}{(\text{SNR}+1)^2}$ bits^2 (a standard formula for the AWGN channel). For $\text{SNR} = 562.3$: $\frac{\text{SNR}(\text{SNR}+2)}{(\text{SNR}+1)^2} = \frac{562.3 \times 564.3}{563.3^2} = \frac{317{,}221}{317{,}107} \approx 1.00036 \approx 1$ (since $\text{SNR} \gg 1$, this ratio $\to 1$). So $V \approx \frac{1}{2}(\log_2 e)^2 = \frac{1}{2} \times (1.4427)^2 = \frac{1}{2} \times 2.0813 = 1.0406^2$.

Rate penalty for $\delta = 10^{-5}$ ($Q^{-1}(10^{-5}) \approx 4.27$): $\sqrt{V/n}\,Q^{-1}(\delta) = \sqrt{1.0406/1.56 \times 10^6} \times 4.27 = \sqrt{6.67 \times 10^{-7}} \times 4.27 = 8.165 \times 10^{-4} \times 4.27 = 0.00349$ bits/use. Relative to $C/W = 9.138$ bits/use, this is a fractional loss of $0.00349/9.138 \approx 0.038\%$ — negligible at this enormous blocklength. This confirms that for deep-space links (long transmission windows, hence huge n), the asymptotic capacity is an excellent approximation, *unlike* the short-packet scenario of Problem 4.4 ($n = 500$, 22% loss).

G.7 End-to-end system view. Putting it together: the 12-bit raw pixel stream (entropy 12 bits/pixel, no structure exploited) is first reduced via predictive transform to a Laplace-distributed residual ($H \approx 6.25$ bits/pixel, Part 1's differential-entropy machinery), then entropy-coded toward this bound via an adaptive arithmetic/Huffman coder (Part 3's source coding theorem, achieving ≈ 6.3 bits/pixel), producing a ≈ 826 KB compressed image, which must be transmitted at ≥ 11.0 kbps. The channel — characterised by its capacity $C(W, \text{SNR})$ (Part 4's Shannon–Hartley / water-filling machinery, here a single band so water-filling is trivial) — requires $W \gtrsim 1.1$–1.3 kHz at the computed link budget, with negligible finite-blocklength penalty (Part 4's dispersion) given the long integration time. Every numbered "Part" of Volume I thus corresponds to one stage of this single pipeline — from raw sensor bits, through compression, to the physical channel.

G.8 Sensitivity analysis: what if the antenna is smaller? Suppose G_T drops from 50 to 40 dBi (a smaller onboard antenna). Then $P_R - N_0 = 58.64 - 10 = 48.64$ dB-Hz, i.e.

linear value 7.31×10^4. Re-solving for W at $R = 11{,}015$ bits/s: try $W = 2000$: SNR $= 7.31 \times 10^4/2000 = 36.55$, $C = 2000 \log_2(37.55) = 2000 \times 5.231 = 10{,}462 < R$. Try $W = 2500$: SNR $= 29.24$, $C = 2500 \log_2(30.24) = 2500 \times 4.918 = 12{,}296 > R$. So $W \approx 2200$–2400 Hz now needed — roughly double the bandwidth for a 10 dB antenna-gain loss, again reflecting the logarithmic capacity-vs-SNR relationship: a 10 dB (factor-of-10) drop in received power requires roughly doubling the bandwidth to maintain the same throughput at this operating point (since $\partial C/\partial W \approx \log_2(\text{SNR})$ when SNR $\gg 1$, and a $10\times$ SNR drop reduces $\log_2(\text{SNR})$ by $\log_2 10 \approx 3.32$ — partially, but not fully, compensated by increasing W, since SNR $= P_R/(N_0 W)$ also decreases as W increases, a second-order effect captured by the transcendental equation solved numerically above rather than by a simple linear rule).

Cross-Volume Connections and Methodological Synthesis

Volume I establishes the foundational vocabulary — entropy, mutual information, KL divergence, the AEP, source coding, and channel capacity — that Volumes II and III build on in two largely independent but complementary directions. This section maps out those connections explicitly, so that a reader who has completed Volume I can navigate the series according to their interests.

To Volume II (Coding Theory & Data Compression). Volume I's Part 3 (source coding, Huffman, Kraft, LZ) is the entry point to Volume II's much deeper treatment:

- **Arithmetic and range coding** generalise the Shannon-code construction of Appendix F.4, removing the $+1$-bit overhead and enabling adaptive, context-dependent models — directly extending the "Computational Notes" sketch of arithmetic coding.
- **Block codes and linear codes** extend the combinatorial bounds of Appendix C (Singleton bound, Hamming bound) into a full algebraic theory (generator matrices, syndrome decoding, the Reed–Solomon construction referenced in the Singleton-bound example).
- **Turbo and LDPC codes** achieve rates approaching the channel capacity C of Part 4 via iterative belief-propagation decoding — an algorithmic descendant of the same "alternating update" philosophy as Blahut–Arimoto (Appendix F.5).
- **Kolmogorov complexity** formalises the "individual sequence" perspective flagged in the Open Problems section, providing the rigorous foundation for universal source coding (LZ, context-tree weighting) beyond the i.i.d. AEP.
- **Rate-distortion theory**, previewed in Appendix E, generalises the Part 3 source coding theorem to *lossy* compression — the rate-distortion function $R(D)$ plays the same role for lossy compression that $H(X)$ plays for lossless compression, and is computed by a Blahut–Arimoto-type algorithm dual to the channel-capacity one (Appendix F.5).

To Volume III (Information Geometry & Statistical Manifolds). Volume I's Part 2 (KL divergence, Bregman divergences in Appendix B) and the maximum-entropy / exponential-family duality (Appendix B) are the seeds of Volume III's geometric perspective:

- **The Fisher information metric** arises as the local (second-order) behaviour of KL divergence: $D_{\mathrm{KL}}(p_\theta \,\|\, p_{\theta+d\theta}) \approx \frac{1}{2} d\theta^\top \mathbf{I}(\theta)\, d\theta$, turning the "global" divergence of Part 2 into a Riemannian metric on the space of distributions.
- **Exponential families**, characterised via convex duality in Appendix B (the log-partition function $A(\theta)$ and its conjugate $A^*(\mu) = -H(\mu)$), become the central geometric objects of Volume III — "flat" with respect to one of the two dual (± 1-)connections induced by the Bregman/KL geometry.
- **The natural gradient**, used throughout modern optimisation (and the companion *Mathematics of Neural Networks* series), is precisely steepest descent with respect to the Fisher-information Riemannian metric — i.e., it re-weights ordinary gradients by the *local KL-divergence geometry* introduced here in its simplest (global, non-differential) form.
- **α-divergences** (of which KL, at $\alpha \to 1$, and the Chernoff information $C(p, q)$ of Appendix D's minimax hypothesis testing, at $\alpha = 1/2$, are special cases) form a one-parameter family interpolating between different geometric structures on the statisti-

cal manifold.

Methodological throughline: convexity. Every major result in Volume I rests on a convexity or concavity statement: Jensen's inequality (Part 1), the log-sum inequality and DPI (Part 2, proven via convexity of KL), the convexity/concavity structure underlying Blahut–Arimoto (Appendix F.5), and the convex duality between entropy and the log-partition function (Appendix B). This is not a coincidence: information-theoretic quantities are, almost without exception, extremal values (maxima of entropy, minima of KL, maxima of mutual information) of convex or concave functionals over convex sets (the simplex Δ_k, spaces of channels, spaces of input distributions) — and the entire apparatus of convex analysis (Legendre–Fenchel duality, Bregman divergences, KKT conditions) is the natural language for these extremal problems. Readers proceeding to Volume III will see this convexity made explicit as the *differential-geometric curvature* of statistical manifolds; readers proceeding to Volume II will see it made explicit as the *algorithmic* convergence guarantees of iterative decoders and compressors.

A unifying numerical thread. The biased four-outcome distribution $p = (0.40, 0.30, 0.20, 0.10)$ introduced in the Part 1 Numerical Tour recurs throughout this volume: its entropy (1.8465 bits) anchors the Huffman code of the Part 3 tour ($L = 1.90$ bits, within 0.0535 of the bound), its joint distribution with Y anchors the mutual-information computation of the Part 2 tour ($I(X;Y) = 0.2631$ bits, verified two independent ways), and its KL-divergence relationships to the uniform distribution echo through Problem 2.1's observation that $D_{\mathrm{KL}}((0.9, 0.1) \,\|\, (0.5, 0.5)) = 0.5310 = C_{\mathrm{BSC}}(0.1)$. A reader wishing to consolidate their understanding of Volume I is encouraged to revisit this single example and verify that every quantity computed from it is mutually consistent — an exercise in "numerical bookkeeping" that mirrors the kind of cross-checking essential in real information-theoretic engineering (as in the deep-space link budget of the Extended Case Study, where source entropy, channel capacity, and dispersion must all be computed from a shared, internally consistent set of numbers).

Bibliography

1. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

2. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, 2006.

3. R. W. Yeung, *Information Theory and Network Coding*, Springer, 2008.

4. I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed., Cambridge University Press, 2011.

5. R. G. Gallager, *Information Theory and Reliable Communication*, Wiley, 1968.

6. A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press, 2011.

7. R. Blahut, "Computation of Channel Capacity and Rate-Distortion Functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

8. S. Arimoto, "An Algorithm for Computing the Capacity of Arbitrary Discrete Memoryless Channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

9. D. A. Huffman, "A Method for the Construction of Minimum-Redundancy Codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 1952.

10. J. Ziv and A. Lempel, "A Universal Algorithm for Sequential Data Compression," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337–343, 1977.

11. I. N. Sanov, "On the Probability of Large Deviations of Random Variables," *Matem. Sbornik*, vol. 42, pp. 11–44, 1957 (English translation: *Selected Translations in Mathematical Statistics and Probability*, vol. 1, 1961).

12. H. Chernoff, "A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the Sum of Observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.

13. Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel Coding Rate in the Finite Blocklength Regime," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

14. D. Slepian and J. K. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

15. R. Ahlswede, "Multi-way Communication Channels," in *Proceedings of the 2nd International Symposium on Information Theory*, Tsahkadsor, Armenia, 1971.

16. P. Bergmans, "Random Coding Theorem for Broadcast Channels with Degraded Components," *IEEE Transactions on Information Theory*, vol. 19, no. 2, pp. 197–207, 1973.

17. R. Etkin, D. Tse, and H. Wang, "Gaussian Interference Channel Capacity to Within One Bit," *IEEE Transactions on Information Theory*, vol. 54, no. 12, pp. 5534–5562, 2008.

18. T. Cover and A. El Gamal, "Capacity Theorems for the Relay Channel," *IEEE Transactions on Information Theory*, vol. 25, no. 5, pp. 572–584, 1979.

19. U. M. Maurer, "Secret Key Agreement by Public Discussion from Common Information," *IEEE Transactions on Information Theory*, vol. 39, no. 3, pp. 733–742, 1993.

20. P. Harremoës and F. Topsøe, "Maximum Entropy Fundamentals," *Entropy*, vol. 3, no. 3, pp. 191–226, 2001.

21. S. Amari, *Information Geometry and Its Applications*, Springer, 2016.

22. M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 4th ed., Springer, 2019.

23. A. N. Kolmogorov, "Three Approaches to the Quantitative Definition of Information," *Problems of Information Transmission*, vol. 1, no. 1, pp. 1–7, 1965.

24. E. T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Review*, vol. 106, no. 4, pp. 620–630, 1957.

25. J. Massey, "Causality, Feedback, and Directed Information," in *Proceedings of the 1990 International Symposium on Information Theory and its Applications*, 1990.

26. N. Merhav, *Statistical Physics and Information Theory*, Foundations and Trends in Communications and Information Theory, vol. 6, nos. 1–2, 2011.

27. A. Lapidoth and S. M. Moser, "Capacity Bounds via Duality with Applications to Multiple-Antenna Systems on Flat-Fading Channels," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2426–2467, 2003.

28. J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed., Cambridge University Press, 2009 (background for the directed-information and causal extensions mentioned in the Open Problems).